PROBABILITY AND STATISTICS

FOR SCIENTISTS AND ENGINEERS

RAO V DUKKIPATI

New Academic Science

YEOVIL COLLEGE
LIBRARY

519
DUK

PROBABILITY AND STATISTICS

FOR SCIENTISTS AND ENGINEERS

YEOVIL COLLEGE
LIBRARY

Yeovil College
Y0069884

Digitized by the Internet Archive
in 2023

https://archive.org/details/probabilitystati0000raov

PROBABILITY AND STATISTICS

FOR SCIENTISTS AND ENGINEERS

RAO V DUKKIPATI
PhD, PE
Fellow of ASME & CSME
Professor, Chair and Graduate Programme
Director, Department of Mechanical Engineering
Fairfield University, Fairfield, Connecticut, USA

NEW ACADEMIC SCIENCE

New Academic Science Limited
The Control Centre, 11 A Little Mount Sion
Tunbridge Wells, Kent TN1 1YS, UK
www.newacademicscience.co.uk
e-mail: info@newacademicscience.co.uk

Copyright © 2011 by New Academic Science Limited
The Control Centre, 11 A Little Mount Sion, Tunbridge Wells, Kent TN1 1YS, UK
www.newacademicscience.co.uk • email: info@newacademicscience.co.uk
Tel: +44(0) 1892 55 7767, Fax: +44(0) 1892 53 0358

ISBN : 978 1 906574 83 3

All rights reserved. No part of this book may be reproduced in any form, by photostat, microfilm, xerography, or any other means, or incorporated into any information retrieval system, electronic or mechanical, without the written permission of the copyright owner.

British Library Cataloguing in Publication Data
A Catalogue record for this book is available from the British Library

Every effort has been made to make the book error free. However, the author and publisher have no warranty of any kind, expressed or implied, with regard to the documentation contained in this book.

Printed and bound in India by Replika Press Pvt. Ltd.

I

Dedicate this book to

Lord Sri Venkateswara

Preface

Engineers and scientists play a significant role in the modern world. They are responsible for research, design and development of diversified products and the manufacturing processes that make these products. Statistics is the branch of applied mathematics that deals with *variability* and its effect on decision making.

Statistics refers to the field or discipline of study. *Statistics* is a group of methods that are used to collect, analyze, present, and interpret data and to make decisions. Decisions made by using statistical methods are called *educated guesses*. *Statistics* has two areas: *theoretical or mathematical* and *applied* statistics. *Theoretical* or *mathematical* statistics deals with the development, deviation, and proof of statistical theorems, formula, rules, and laws. *Applied statistics* concerns the applications of those theorems, formulas, rules and laws to solve real-world problems. This book deals with *applied statistics*.

Applied statistics is divided into two branches: *descriptive statistics* and *inferential statistics*. *Descriptive statistics* involves methods for organizing, displaying, and describing data by using tables, graphs and summary measures. The collection of all elements of interest is called a *population* in statistics. The selection of a few elements from this population is called a *sample*. *Statistics* deals with making decisions, inferences, predictions, and forecasts about populations based on results obtained from samples. Statistics that deals with decision-making procedures is called *inferential statistics*. This branch of statistics is also known as *inductive reasoning or inductive statistics*. *Inferential statistics* consists of methods that employ sample results to help make decisions or predictions about a population.

Probability, which measures the likelihood that an event will occur, is an important part of statistics. It is the basis of inferential statistics, where decisions are made under conditions of uncertainty. Probability theory is used to evaluate the uncertainty involved in those decisions. Probability statements are about occurrence or non-occurrence of a certain event under certain conditions.

Each chapter in this book consists of a concise but thorough fundamental statement of the theory, principles and methods, followed by numerous illustrative worked examples. A large number of sample unworked problems for students' practice, to amplify and extend the theory are also included at the end of each chapter.

The course material presented in the book has evolved from the author's experience in teaching this course to the undergraduate and graduate students of science and engineering at universities in Canada, US, and the industry during the last 35 years. Probability and Statistics is intended as a textbook for undergraduate and graduate students in mathematics, science and engineering curriculum. This book also serves a sound basic background for solving many practical probability and statistics problems or for pursuing more advanced or specialized studies.

I sincerely hope that the final outcome of this book helps the students in developing an appreciation for the topics of probability and statistics. Answers to all the end of chapter problems are given in the book.

Rao V. Dukkipati

Acknowledgements

I am grateful to all those who have had a direct impact on this work. Many people working in the general area of probability and statistics have influenced the format of this book. I would very much like to acknowledge the encouragement, patience and support provided to me by my family members: my wife, Sudha, my family members, Ravi, Madhavi, Anand, Ashwin, Raghav, and Vishwa who have also shared in all the pain, frustration, and fun of producing a manuscript.

I would appreciate being informed of errors, or receiving other comments about the book.

Rao V. Dukkipati

Contents

Preface *vii*
Acknowledgements *ix*

1. NUMERICAL DESCRIPTIVE MEASURES **1–35**

1.1 Introduction *1*
  1.1.1 Population and Sample *2*
  1.1.2 Types of Variables *2*
  1.1.3 Organising Data *3*
    1.1.3.1 Qualitative Data *4*
    1.1.3.2 Graphical Representation of Qualitative Data *4*
    1.1.3.3 Graphical Representation of Quantitative Data *5*
1.2 Numerical Summary Measures *9*
  1.2.1 Measures of Central Tendency for Ungrouped Data *9*
    1.2.1.1 Mean for Ungrouped Data *10*
    1.2.1.2 Median *10*
    1.2.1.3 Mode *10*
    1.2.1.4 Empirical Relation among Mean, Median and Mode *10*
  1.2.2 Measures of Dispersion for Ungrouped Data *12*
    1.2.2.1 Range *12*
    1.2.2.2 Variance and Standard Deviation *12*
  1.2.3 Mean, Variance and Standard Deviation for Grouped Data *14*
    1.2.3.1 Mean for Grouped Data *14*
    1.2.3.2 Variance and Standard Deviation for Grouped Data *14*
  1.2.4 Measures of Position *15*
    1.2.4.1 Quartiles and Interquartile Range *16*
    1.2.4.2 Percentiles *17*

1.2.4.3 Skewness and Kurtosis *18*
1.2.4.4 Box-and-Whisker Plot *19*
1.3 Summary *21*
Problems 22
Review Questions 31
State True or False 32
Answers to State True or False 35

2. PROBABILITY 36–78

2.1 Experiment, Outcome and Sample Space *36*
2.2 Simple and Composite Events *37*
2.3 Axioms of Probability *38*
2.4 Finite Probability Spaces *38*
2.5 Infinite Probability Spaces *39*
2.6 Properties of Probability *40*
2.7 Venn Diagram *40*
2.8 Probability Tree or Tree Diagram *41*
2.9 Approaches to Probability *44*
2.9.1 Classical Probability *45*
2.9.2 Relative Frequency Concept of Probability *45*
2.9.3 Subjective Probability *45*
2.9.4 Marginal Probability *45*
2.9.5 Conditional Probability *45*
2.10 Mutually Exclusive Events *47*
2.11 Independent and Dependent Events *48*
2.12 Complementary Events *48*
2.13 Intersection of Events and Multiplication Rule *51*
2.13.1 Intersection of Events *51*
2.13.2 Multiplication Rule *51*
2.14 Union of Events and the Addition Rule *55*
2.14.1 Union of Events *55*
2.14.2 Addition Rule *55*
2.15 Baye's Formula *58*
2.16 Additional Examples and Solutions *61*
2.17 Summary *67*
Problems 67
Review Questions 76
State True or False 77
Answers to State True or False 78

3. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS 79–173

3.1 Random Variables *79*
  3.1.1 Discrete Random Variables *80*
  3.1.2 Mean and Standard Deviation of a Discrete Random Variable *82*
  3.1.3 Continuous Random Variables *90*
  3.1.4 Mean and Variance for Continuous Random Variables *91*
  3.1.5 Expectation *92*
3.2 Permutations and Combinations *98*
  3.2.1 Permutations *99*
  3.2.2 Combinations *101*
3.3 Discrete Distributions *103*
  3.3.1 Hypergeometric Distribution *103*
  3.3.2 The Binomial Probability Distribution *108*
  3.3.3 The Binomial Experiment *108*
  3.3.4 The Binomial Formula *108*
    3.3.4.1 Binomial Theorem *108*
    3.3.4.2 Cumulative Terms for Binomial Distribution *109*
    3.3.4.3 Mean and Standard Deviation of Binomial Distribution *110*
  3.3.5 Poisson Distribution *118*
    3.3.5.1 Derivation from Binomial Distribution *120*
    3.3.5.2 Mean and Standard Deviation *120*
3.4 Continuous Probability Distributions *127*
  3.4.1 The Normal Distribution *127*
    3.4.1.1 Properties of the Normal Distribution *129*
    3.4.1.2 Mean and Variance of the Normal Distribution *129*
    3.4.1.3 The Cumulative Normal Distribution *130*
    3.4.1.4 The Standard Normal Distribution *130*
    3.4.1.5 Problem-Solving Procedure *132*
3.5 Approximating Probability Distributions *145*
  3.5.1 Binomial Approximation to the Hypergeometric *145*
  3.5.2 Poisson Approximation to the Binomial *145*
  3.5.3 Normal Approximation to the Binomial *146*
  3.5.4 Normal Approximation to the Poisson *147*
3.6 Chebyshev's Theorem *152*
3.7 Empirical Rule *156*
3.8 The Central Limit Theorem *157*
Problems 159
Review Questions 170
State True or False 171
Answers to State True or False 173

4. SAMPLING DISTRIBUTIONS 174–203

4.1 Properties of Sample Mean and Variance *175*
4.2 Population and Sampling Distributions *176*
  4.2.1 Population Distribution *176*
  4.2.2 Sampling Distribution *176*
4.3 Sampling and Nonsampling Errors *177*
4.4 Mean and Standard Deviation of x *178*
4.5 Shape of the Sampling Distribution of $\overline{x}$ *182*
  4.5.1 Sampling from a Normally Distributed Population *182*
  4.5.2 Sampling from a Population that is not Normally Distributed *184*
4.6 Applications of the Sampling Distribution of $\overline{x}$ *184*
4.7 Population and Sample Proportions *185*
4.8 Sampling Distribution of $\hat{p}$ *186*
4.9 Mean and Standard Deviation of $\hat{p}$ *186*
4.10 The Chi-square Distribution *189*
4.11 The t-Distribution *192*
4.12 The F-Distribution *194*
4.13 Summary *197*
Problems *198*
Review Questions *201*
State True or False *202*
Answers to State True or False *203*

5. ESTIMATION 204–256

5.1 Point Estimation *204*
5.2 Interval Estimation *205*
5.3 Confidence Interval on Mean, Variance Known *207*
5.4 Confidence Interval on the Mean of a Normal Distribution, Variance Unknown *214*
5.5 Confidence Interval on the Variance of a Normal Distribution *218*
5.6 Confidence Interval on a Population Proportion *220*
5.7 Confidence Interval on the Difference in Two Means, Variance Known *224*
5.8 Confidence Interval on the Difference in Means of Two Normal Distributions, Variances Unknown *226*
5.9 Confidence Interval on $\mu_1 - \mu_2$ for Paired Observations *233*
5.10 Confidence Interval on the Ratio of Variance of Two Normal Distributions *236*
5.11 Confidence Interval on the Difference in Two Proportions *239*
5.12 Sample Size Selection *244*
  5.12.1 Sample Size Selection for Estimating Population Mean *244*
  5.12.2 Sample Size for the Estimation of Proportion *245*

5.13 Summary *246*
Problems *246*
Review Questions *255*
State True or False *255*
Answers to State True or False *256*

6. HYPOTHESIS TESTING 257–323

6.1 Null Hypothesis and Alternative Hypothesis *257*
6.2 The Critical Region *259*
6.3 Types of Sampling Error (Type I and Type II Errors) *259*
6.4 Level of Significance *260*
6.5 Tails of a Test *261*
6.6 Hypothesis Test on the Population Mean, Standard Deviation Known *262*
6.7 Hypothesis Test on the Population Mean, Standard Deviation Unknown *267*
6.8 Hypothesis Test for a Population Variance *272*
6.9 Hypothesis Test on a Population Proportion *276*
6.10 Hypothesis Test on Equality of Two Means, Variances Known *280*
6.11 Hypothesis Test on the Means of Two Normal Distributions, Variances Unknown *285*
  6.11.1 Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ *285*
  6.11.2 Case 2: $\sigma_1^2 \neq \sigma_2^2$ *290*
6.12 Hypothesis Test to Compare Two Population Means (Paired *t*-test) *295*
6.13 Hypothesis Test on the Equality of Two Variances *300*
6.14 Hypothesis Test on Two Proportions *305*
6.15 Summary *312*
Problems *313*
Review Questions *320*
State True or False *321*
Answers to State True or False *323*

7. CURVE FITTING, REGRESSION AND CORRELATION 324–380

7.1 Linear Equation *325*
7.2 Curve Fitting with a Linear Equation *326*
7.3 Criteria for a "Best" Fit *328*
7.4 Linear Least-Squares Regression *329*
7.5 Linear Regression Analysis *331*
7.6 Interpretation of *a* and *b* *334*
7.7 Standard Deviation of Random Errors *335*
7.8 Coefficient of Determination *335*
7.9 Sampling Distribution of *b* *337*
7.10 Hypothesis Testing about *B* *339*
7.11 Linear Correlation *340*

7.12 Estimating the Mean Value of y *344*
7.13 Estimating a Particular Value of y *345*
7.14 Linearization of Nonlinear Relationships *347*
7.15 Polynomial Regression *350*
7.16 Quantification of Error of Linear Regression *353*
7.17 Multiple Linear Regression *355*
7.18 Weighted Least-Squares Method *357*
7.19 Orthogonal Polynomials and Least Squares Approximation *358*
7.20 Least-Squares Method for Continuous Data *358*
7.21 Approximation using Orthogonal Polynomials *359*
7.22 Gram-Schmidt Orthogonalization Process *361*
7.23 Additional Example Problems and Solutions *362*
7.24 Summary *367*
Problems 367
Review Questions 377
State True or False 379
Answers to State True or False 380

8. CHI-SQUARE TESTS 381–400

8.1 Introduction *381*
8.2 A Goodness-of-Fit Test *381*
8.3 Contingency Table *385*
8.4 A Test of Independence or Homogeneity *385*
8.5 Inferences about the Population Variance *391*
8.6 Estimation of the Population Variance *391*
8.7 Hypothesis Tests about the Population Variance *393*
Problems 395
Review Questions 399
State True or False 400
Answers to State True or False 400

9. ANALYSIS OF VARIANCE 401–411

9.1 Introduction *401*
9.2 One-way Analysis of Variance *401*
9.3 Test Statistic *402*
9.4 One-way ANOVA Test *404*
Problems 407
Review Questions 409
State True or False 410
Answers to State True or False 411

APPENDICES 412–442

Appendix–A Values for *n* Factorial *412*
Appendix–B Binomial Coefficients *413*
Appendix–C Table of Binomial Probabilities *414*
Appendix–D Cumulative Poisson Distibutions *422*
Appendix–E Cumulative Standard Normal Distribution *428*
Appendix–F chi-square Distribution Table *430*
Appendix–G The *t*-Distribution Table *431*
Appendix–H The *F*-Distribution Table *434*
Appendix–I Cramer's Rule *439*

BIBLIOGRAPHY 443–445

GLOSSARY OF TECHNICAL TERMS 446–460

GLOSSARY OF SYMBOLS 461–464

KEY FORMULAE 465–477

ANSWERS TO SELECTED PROBLEMS 478–513

CHAPTER 1

Numerical Descriptive Measures

This chapter introduces the areas of statistics, basic concepts and definitions, population and sample, types of variables, graphical representations of qualitative and quantitative data, numerical summary measures, measure of central tendency for ungrouped data (mean, median and mode), measures of dispersion for ungrouped data (range, variance and standard deviation), mean, variance and standard deviation for grouped data, measures of position including quartiles, percentiles, box-and-whisker plot, skewness and kurtosis.

1.1 INTRODUCTION

Statistics refers to the field or discipline of study. *Statistics* is a group of methods that are used to collect, analyse, present, and interpret data and to make decisions. Decisions made by using statistical methods are called *educated guesses*. *Statistics* has two areas, *theoretical* or *mathematical statistics* and *applied statistics* as shown in Fig.1.1. *Theoretical* or *mathematical statistics* deals with the development, derivation, and proof of statistical theorems, formula, rules, and laws. *Applied statistics* concerns the applications of those theorems, formulae, rules and laws to solve real-world problems. This chapter deals with *applied statistics.*

Applied statistics is divided into two branches, *Descriptive statistics* and *Inferential statistics*. *Descriptive statistics* involves methods for organising, displaying, and describing data by using tables, graphs and summary measures. The collection of all elements of interest is called a *population* in statistics. The selection of a few elements from this population is called a *sample*. Statistics that deals with making-decision, inferences, predictions, and forecasts about populations based on results obtained from samples. Statistics that deals with decision-making procedures is called *inferential statistics*. This branch of statistics is also known as *inductive reasoning* or *inductive statistics*. *Inferential statistics* consists of methods that employ sample results to help make decisions or predictions about a population.

Probability, which measures the likelihood that an event will occur, is an important part of statistics. It is the basis of inferential statistics, where decisions are made under conditions of uncertainty. Probability theory is used to evaluate the uncertainty involved in those decisions. Probability statements are about occurrence or non-occurrence of a certain event under certain conditions.

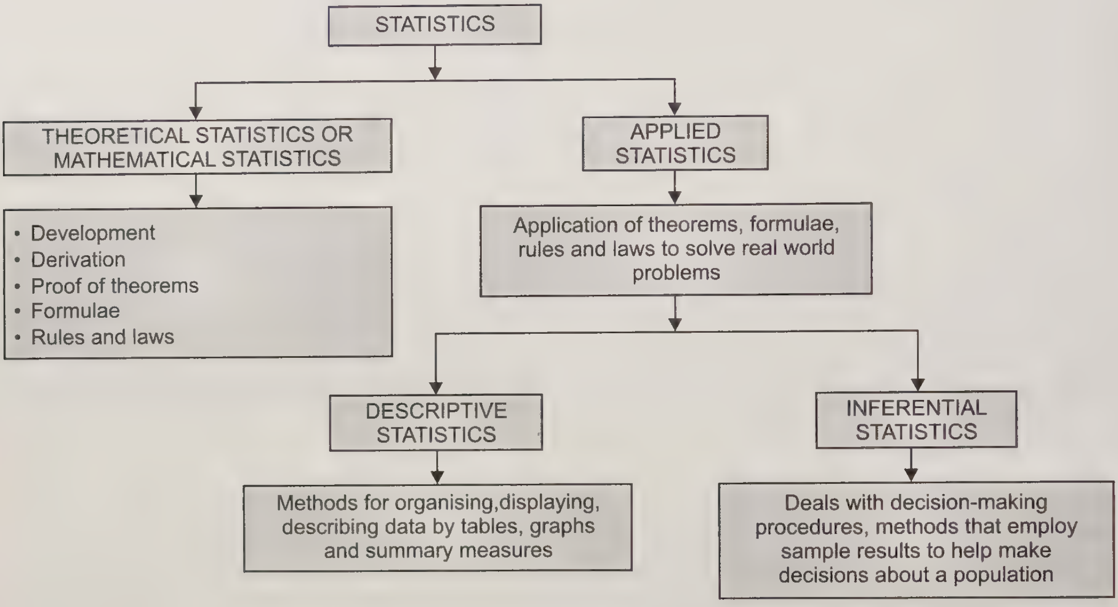

STATISTICS
THEORETICAL STATISTICS OR MATHEMATICAL STATISTICS
APPLIED STATISTICS
• Development
• Derivation
• Proof of theorems
• Formulae
• Rules and laws
Application of theorems, formulae, rules and laws to solve real world problems
DESCRIPTIVE STATISTICS
INFERENTIAL STATISTICS
Methods for organising,displaying, describing data by tables, graphs and summary measures
Deals with decision-making procedures, methods that employ sample results to help make decisions about a population

Fig. 1.1: Areas of statistics

1.1.1 Population and Sample

A *population* consists of all elements: individuals, items, or objects, whose characteristics are being studied. The population that is being studied is also called the *target population*. The collection of a few elements selected from a population is called a sample. Thus, a portion of the population selected for study is referred to as a *sample*. The collection of information from the elements of a population or a sample is called a survey. A *survey* that includes each and every element of the population is called a *census*. A survey conducted on a sample is called a *sample survey*. A sample that represents the characteristics of the population as closely as possible is called a *representative sample*.

A sample drawn in such a way that each element of the population has same chance of being selected is called a *random sample*. If the chance of being selected is the same for each element of the population, it is called a *simple random sample*. An *element* or *member* of a sample or population is a specific subject or object about which the information is collected. A *variable* is a characteristic under study that assumes different values for different elements. In relation to a variable, the value of a *constant* is fixed. The value of a variable for an element is called an *observation* or *measurement*. A *data set* is a collection of observations on one or more variables.

1.1.2 Types of Variables

A variable that can be measured numerically is called a *quantitative variable*. The data collected on a quantitative variable are known as *quantitative data*. Quantitative variable is classified as either *discrete variable* or *continuous variable*. A variable whose values are continuous is called a *continuous variable*. A *discrete variable* can assume only certain values with no intermediate values. On the other hand, a variable that can assume any numerical value over a certain interval or intervals is called a *continuous variable*. The time taken to complete a trip to a corner store is an example of a continuous random variable because it can assume any value, say, between one and two hours. A variable that cannot assume a numerical value but can be classified into two or more non-numeric categories is called a *qualitative* or *categorical* variable. The data collected on such a variable are known as *qualitative data*. The types of variables and types of data are classified as shown in Figs.1.2 *(a)* and *(b)*.

Quantitative

Continuous: Weight, Height, Line width measurements

Discrete: Number of students, Number of accidents

Qualitative

Nominal: Black, Green, Yellow, ...

Ordinal: High, Medium, Low

Ranked: 1 2 3 etc.

Data collected on different elements at the same point in time or for the same period of time are called *cross-section data.*

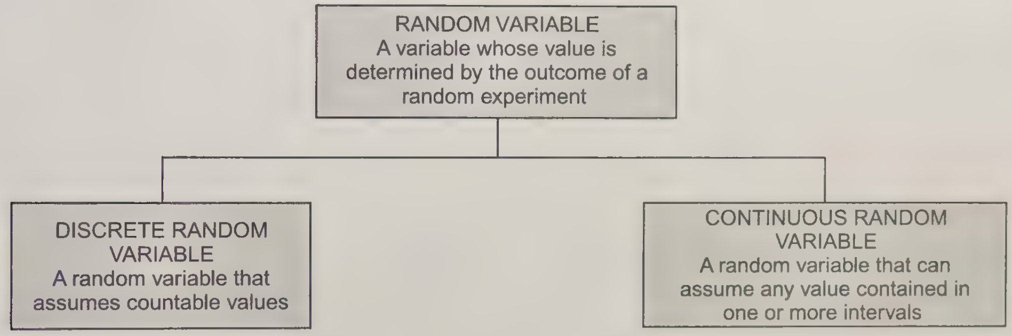

RANDOM VARIABLE
A variable whose value is determined by the outcome of a random experiment
DISCRETE RANDOM VARIABLE
A random variable that assumes countable values
CONTINUOUS RANDOM VARIABLE
A random variable that can assume any value contained in one or more intervals

Fig. 1.2(a): Types of variables

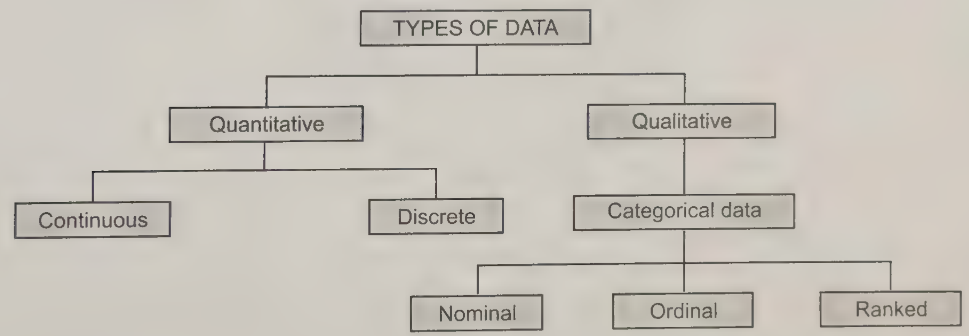


Fig. 1.2(b): Types of data

1.1.3 Organising Data

Data recorded in the sequence in which they are collected and before they are processed or ranked are called *raw data*.

1.1.3.1 *Qualitative Data*

A frequency *distribution* for qualitative data lists all categories and the number of elements that belong to each of the categories. A *relative frequency distribution* lists the relative frequencies for all categories.

$$\text{Relative frequency of a category} = \frac{\text{frequency of that category}}{\text{sum of all frequencies}} \tag{1.1}$$

A *percentage distribution* lists the percentages for all categories.

$$\text{Percentage} = \text{relative frequency} \times 100 \tag{1.2}$$

1.1.3.2 *Graphical Representation of Qualitative Data*

The *bar graph* and the *pie chart* are two types of graphs used to represent or display qualitative data. A graph made of bars whose heights represent the frequencies of representative categories is called a *bar graph*. A *pie chart* is more commonly used to display percentages, although it can be used to display frequencies or relative frequencies. The pie chart represents the total sample or population. The pie chart is divided into different portions that represent the percentages of the population or sample belonging to different categories.

Example E 1.1

The following data give the results of a sample survey of three categories A, B and C.

A B C C A B A B C C A B C A C

A C A B C A B C C C C A C C C

(a) prepare a frequency distribution table

(b) compute the relative frequencies and percentages for all three categories A, B and C

(c) what percentage of the elements in this sample belongs to category B?

(d) what percentage of the elements in this sample belongs to categories A or C?

(e) draw a bar graph of the frequency distribution.

SOLUTION:

(a) and *(b)*

Category	Frequency	Relative frequency	Percentage
A	9	0.30	30
B	6	0.20	20
C	15	0.50	50
Totals	30	1.00	100

(c) 20% of the element in this sample belongs to category B.

(d) 30% + 50% = 80% of the elements in this sample belong to categories A or C.

(e) The bar graph of the frequency distribution is shown in Fig. E1.1.

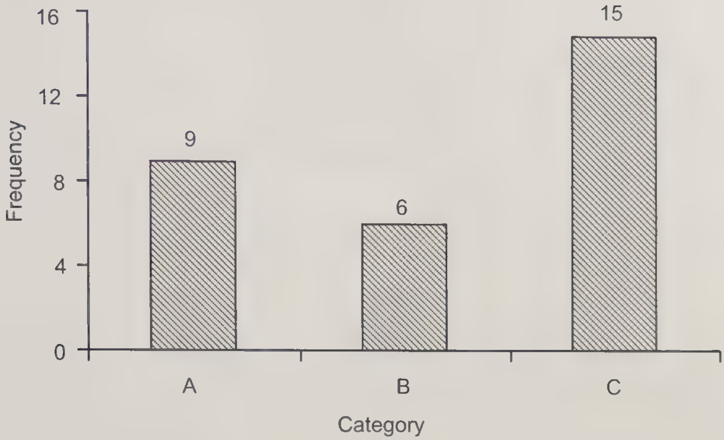


Fig. E1.1: Bar graph of the frequency distribution

Example E1.2

Draw the pie chart for the percentage distribution in Example E1.1.

Solution:

The pie chart is shown in Fig. E1.2.

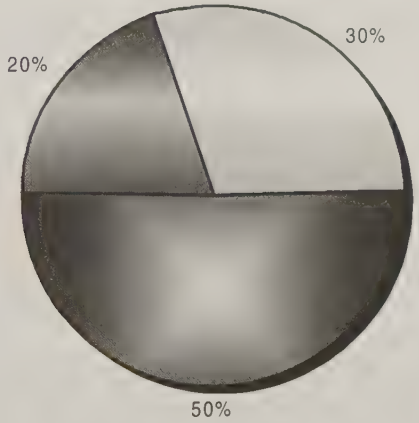


Fig. E1.2: Pie chart

1.1.3.3 Graphical Representation of Quantitative Data

A *frequency distribution* for quantitative data lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called *grouped data*. The *class boundary* is the midpoint of the upper limit of one class and the lower limit of the next class, while class width = upper boundary – lower boundary.

The *class width* is also known as the *class size*.

$$\text{The } \textit{class} \text{ midpoint or } \textit{mark} = \frac{\text{lower limit} + \text{upper limit}}{2} \tag{1.3}$$

$$\text{Approximate class width} = \frac{\text{largest value} - \text{smallest value}}{\text{number of classes}} \tag{1.4}$$

The relative frequencies and percentages for a quantitative data set are calculated as follows:

$$\text{Relative frequency of a class} = \frac{\text{frequency of that class}}{\text{sum of all frequencies}} = \frac{f}{\sum f} \tag{1.5}$$

$$\text{Percentage} = \text{relative frequency} \times 100$$

A *histogram* is a graph in which classes are marked on the horizontal axis and the frequencies, relative frequencies or percentages are marked on the vertical axis. The frequencies, relative frequencies, or percentage are represented by the heights of the bars. In a histogram, the bars are drawn adjacent to each other.

The most common shapes of histograms are *symmetric, skewed,* and *uniform* or *rectangular.* A *symmetric histogram* is identical on both sides of its central point. A *skewed histogram* is nonsymmetric. For a skewed histogram, the tail on one side is longer than the tail on the other side. A *skewed-to-the-right* histogram has a longer tail on the right side while a skewed-to-the-left histogram has a longer tail on the left side. A *uniform* or *rectangular histogram* has the same frequency for each class.

A *polygon* is another method that can be used to represent quantitative data in graphical form. A graph formed by joining the midpoints of the tops of successive bars in a histogram with straight lines is called a *polygon.* A polygon with relative frequencies marked on the vertical axis is called a *relative frequency polygon.* In a similar manner, a polygon with percentages marked on the vertical axis is called a *percentage polygon.*

A *cumulative frequency distribution* gives the total number of values that fall below the upper boundary of each class.

$$\text{Cumulative relative frequency} = \frac{\text{cumulative frequency}}{\text{total observations in the data set}} \tag{1.6}$$

$$\text{Cumulative frequency} = \text{cumulative relative frequency} \times 100 \tag{1.7}$$

When plotted on a graph, the cumulative frequencies give a curve that is called an *ogive.* Thus, an *ogive* is a curve drawn for the cumulative frequency distribution by joining with straight lines the dots marked above the upper boundaries of classes at height equal to the cumulative frequencies of respective classes.

Another technique that is used to represent quantitative data in condensed form is the *stem-and-leaf-display.* In a *stem-and-leaf display* of quantitative data, each value is divided into two portions: a stem and a leaf. The leaves for each stem are shown separately in a display.

Example E1.3

Out of 27 randomly selected families in a small town, the following data give the number of children less than 18 years of age:

1	0	2	4	4	2	3	2	0
2	0	0	0	0	1	4	0	2
0	2	0	2	3	2	1	2	2

(*a*) construct a frequency distribution table using single-valued classes

(*b*) compute the relative frequencies and percentages for all classes

(*c*) find the number of families with one or two children less than 18 years of age

(*d*) find the number of families with two or three children less than 18 years of age

(*e*) draw a bar graph for the frequency distribution.

SOLUTION:

(*a*) and (*b*) The frequency distribution, relative frequencies and percentages are shown in Table E1.3.

Table E1.3

Number of children less than 18 years of age	Frequency	Relative frequency	Cumulative relative frequency	Percentage (probability)
0	9	0.334	0.334	33.3
1	3	0.111	0.445	11.1
2	10	0.370	0.815	37.0
3	2	0.074	0.889	7.4
4	3	0.111	1.0	11.1
Total	27			100.0

(*c*) Number of families with one or two children under 18 years of age = 3 + 10 = 13.

(*d*) Number of families with two or three children under 18 years of age = 10 + 2 = 12.

(*e*) The bar graph for the frequency distribution is shown in Fig. E1.3.

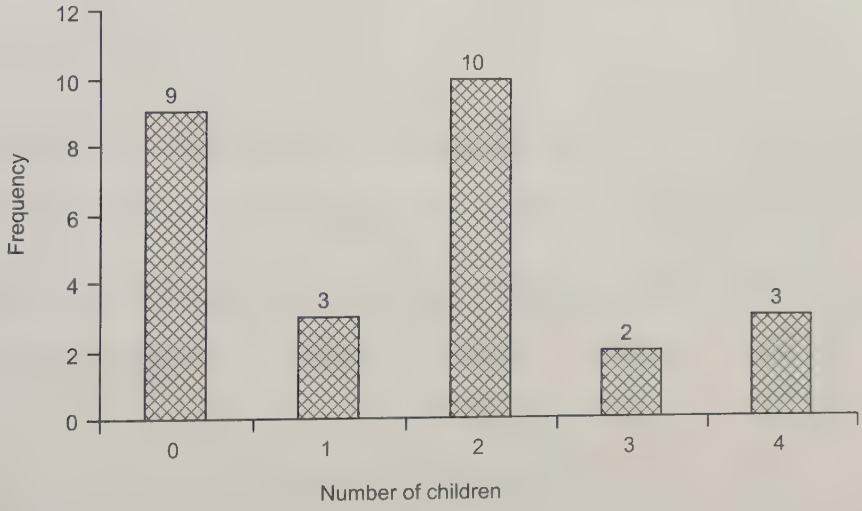


Fig. E1.3: Bar graph of frequency distribution

Example E1.4

The frequency distribution of ages of all 60 part-time graduate students in a Master of Science in Mechanical Engineering degree program of a university is given below:

Age (years)	Number of students
21-25	22
26-30	13
31-35	11
36-40	8
41-45	6

(*a*) prepare a cumulative frequency distribution table
(*b*) calculate the cumulative relative frequencies and cumulative percentages for all classes
(*c*) find the percentage of students who are 31 years of age or older
(*d*) draw a ogive for the cumulative percentage distribution
(*e*) using the ogive, determine the percentage of students who are 35 years or younger.

SOLUTION:

(*a*) and (*b*) The cumulative frequency, cumulative relative frequency and percentages are shown in Table E1.4.

Table E1.4

Age (years)	Frequency	Cumulative frequency	Cumulative relative frequency	Cumulative percentage
21–25	22	22	0.366	36.6
26–30	13	35	0.583	58.3
31–35	11	46	0.766	76.6
36–40	8	54	0.90	90.0
41–45	6	60	1.00	100.0

(*c*) Percentage of students who are 31 years of age or older = 100 – 58.3 = 41.7%.
(*d*) The frequency, cumulative relative frequency and the ogive diagram for the cumulative percentage distribution is shown in Fig. E1.4.
(*e*) The percentage of a student who is 35 years or younger is about 70%.

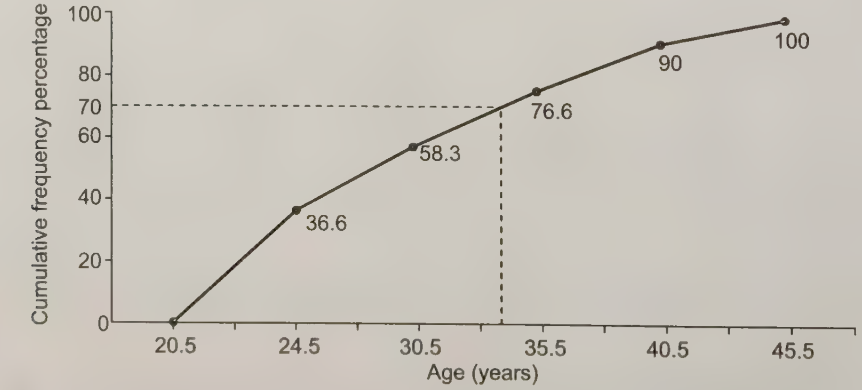


Fig. E1.4: Ogive diagram for the cumulative percentage distribution

Example E1.5

Following are the overtime monies earned by 15 employers of a company during the month of August 2008.

780, 892, 995, 1168, 883, 1041, 994, 1084, 776, 981, 1129, 1067, 932, 655, 642

Prepare a stem-and-leaf display by arranging the leaves for each stem in an increasing order.

SOLUTION:

The stem-and-leaf display and the stem-and-leaf display data are shown in Figs. E1.5(*a*) and E1.5(*b*) respectively.

6	55	42		
7	80	76		
8	92	83		
9	95	94	81	32
10	41	84	67	
11	68	29		

Fig. E1.5(a): Stem-and-leaf display of data

6	42	55		
7	76	80		
8	83	92		
9	32	81	94	95
10	41	67	84	
11	29	68		

Fig. E1.5(b): Stem-and-leaf display of data arranged in increasing order

1.2 NUMERICAL SUMMARY MEASURES

There are three basic numerical summary (descriptive) measures used to organize and display large data sets as shown in Fig.1.3. They are:

1. Measures of central tendency
2. Measures of dispersion or scatter and
3. Measures of position.

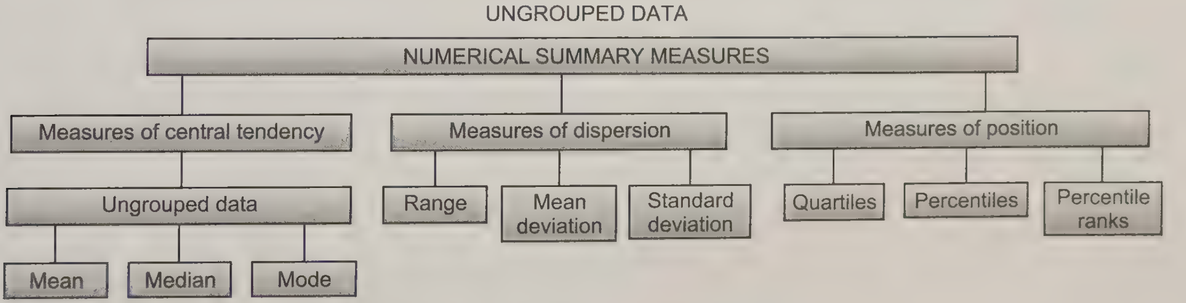


Fig. 1.3: Three basic numerical summary measures

1.2.1 Measures of Central Tendency for Ungrouped Data

A data set is generally represented by numerical summary measures called the *typical values*. A measure of central tendency gives the center of a histogram or a frequency distribution curve. There are three measures of central tendency: the mean, the median, and the mode.

1.2.1.1 Mean for Ungrouped Data

An *average* is a value which is typical or representative of a set of data. Since such values tend to lie centrally within a set of data arranged according to magnitude, averages are also called *measures of central tendency.*

The mean, also called the arithmetic mean or average, for ungrouped data is obtained by dividing the sum of all values by the number of values in the data set. Thus,

$$\text{Mean for population data } \mu = \frac{\sum X}{N} \tag{1.8}$$

$$\text{Mean for sample data } \bar{X} = \frac{\sum X}{n} \tag{1.9}$$

where μ is the population mean, $\sum X$ the sum of all values, N the population size, $\bar{X}$ the sample mean, and n the sample size. Sometimes the data may contain a few very small or a few very large values. Such values are called *outliers* or *extreme values.*

1.2.1.2 Median

The *median* is the value of the middle term in a data set that has been ranked in either increasing or decreasing order. Median for ungrouped data = the value of the $\left(\frac{n+1}{2}\right)^{th}$ term in a ranked data set. If the given data set represents a population, replace n by N. If the number of observations in a data set is odd, then the median is given by the value of the middle term in the ranked data. Similarly, if the number of observations is even, then the median is the average of the values of the two middle terms.

1.2.1.3 Mode

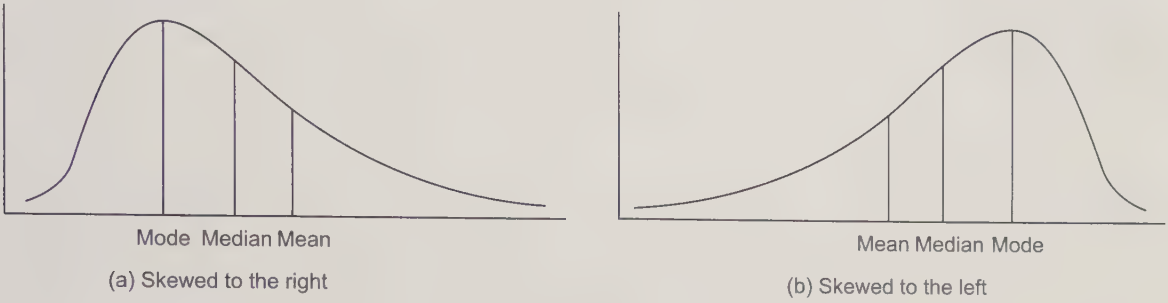
The *mode* for an ungrouped data is the value that occurs with the highest frequency in a data set. A data set with only one value occurring with highest frequency has only one mode and such a data set is called *unimodal.* A data set with two values occurring with the same (highest) frequency has two modes. The distribution in this case is called *bimodal.* If more than two values in a data set occur with the same (highest) frequency, then the data set contains more than two modes and it is said to be *multi modal.* It should be noted here that a data set with each value occurring only once has no mode.

1.2.1.4 Empirical Relation among Mean, Median and Mode

For unimodal frequency curves which are moderately skewed (asymmetrical), the empirical relationship is given by

$$\text{mean} - \text{mode} = 3\ (\text{mean} - \text{median})$$

Figures 1.4*(a)* and *(b)* show the relative positions of the mean, median and mode for frequency curves which are skewed to the right and left respectively. It should be noted here that for symmetrical curves, the mean, median and mode and they all coincide.



(a) Skewed to the right

(b) Skewed to the left

Fig. 1.4: Empirical relationship

Example E1.6

The following data gives the number of automobiles owned by each of the 30 families in a city.

2	3	1	1	1	2	4	3	1	1
3	2	2	1	2	3	2	3	1	3
4	1	3	4	1	1	1	3	2	2

(a) prepare a frequency distribution table for this data set using single-valued classes.

(b) calculate the relative frequency and percentage distributions.

(c) find the percentage of the families in this sample having two or three automobiles.

(d) draw a bar graph for the relative frequency distribution.

SOLUTION:

(a) and *(b)*

Number of automobiles in a family	Frequency	Relative frequency	Percentage
1	11	0.366	36.6
2	8	0.267	26.7
3	8	0.267	26.7
4	3	0.1	10

(c) Number of families in this sample who have two or three automobiles = 26.7 + 26.7 = 53.4%

(d) The bar graph for the relative frequency distribution is shown in Fig. E1.6.

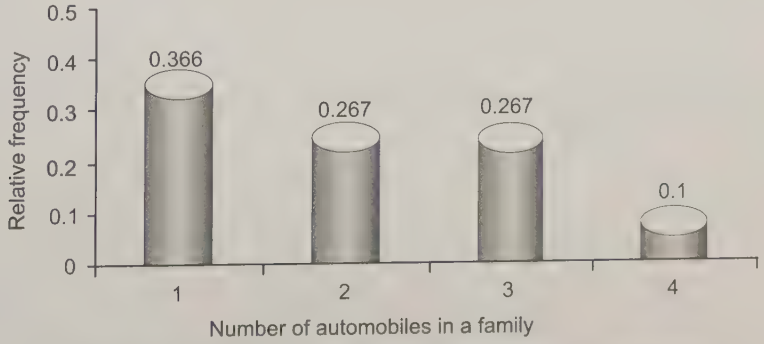


Fig. E1.6: Bar graph of relative frequency distribution

Example E1.7

The following are the ages of 15 employees of a manufacturing company.

65, 38, 52, 27, 24, 45, 49, 50, 26, 37, 44, 52, 41, 60, 39

Calculate the mean, median and mode.

SOLUTION:

$$\text{Mean} = \frac{\Sigma X}{n} = \frac{(65+38+52+27+24+45+49+50+26+37+44+52+41+60+39)}{15}$$

$$= \frac{649}{15} = 43.266 \text{ years.}$$

Rearrange the given data in an increasing order gives

24, 26, 27, 37, 38, 39, 41, 44, 45, 49, 50, 52, 52, 60, 65

$$\text{Position of the middle term} = \frac{n+1}{2} = \frac{15+1}{2} = 8$$

Hence, the median is the value of the middle term in the ranked data.

Median = 44 years

In the data set, 52 occur twice and each of the remaining values occurs only once.

Therefore, mode = 52.

1.2.2 Measures of Dispersion for Ungrouped Data

The mean, median and mode are not sufficient measures to reveal the shape of the distribution of a data set. The measures that show the spread of a data set are called the *measures of dispersion*. There are three measures of dispersion range, variance, and standard deviation as shown in Fig. 1.5.

1.2.2.1 Range

The range is the difference between the largest and smallest values in the data.

The range for ungrouped data = largest value – smallest value.

The range is not a very satisfactory measure of dispersion.

1.2.2.2 Variance and Standard Deviation

The standard deviation is the most used measure of dispersion. The formulae for calculating the variance and standard deviation are as follows:

Variance for ungrouped data

Population variance

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N} = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{N}}{N} \tag{1.10}$$

Sample variance

$$s^2 = \frac{\sum\left(x-\overline{X}\right)^2}{n-1} = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n-1} \tag{1.11}$$

where, σ^2 is the population variance, s^2 is the sample variance, $(x - \mu)$ or $(x - \bar{x})$ is the deviation of x value from the mean, μ is the population mean, $\bar{x}$ is the sample mean, N is the population size, and n is the sample size.

The population standard deviation $\sigma = \sqrt{\text{population variance}} = \sqrt{\sigma^2}$ (1.12)

The sample standard deviation $s = \sqrt{\text{sample variance}} = \sqrt{s^2}$ (1.13)

A numerical value such as the mean, median, mode, range, variance, or standard deviation computed for a population data set is called a *population parameter*, or simply a *parameter*. A summary measure calculated for a sample data set is called a *sample statistic* or simply a *statistic*.

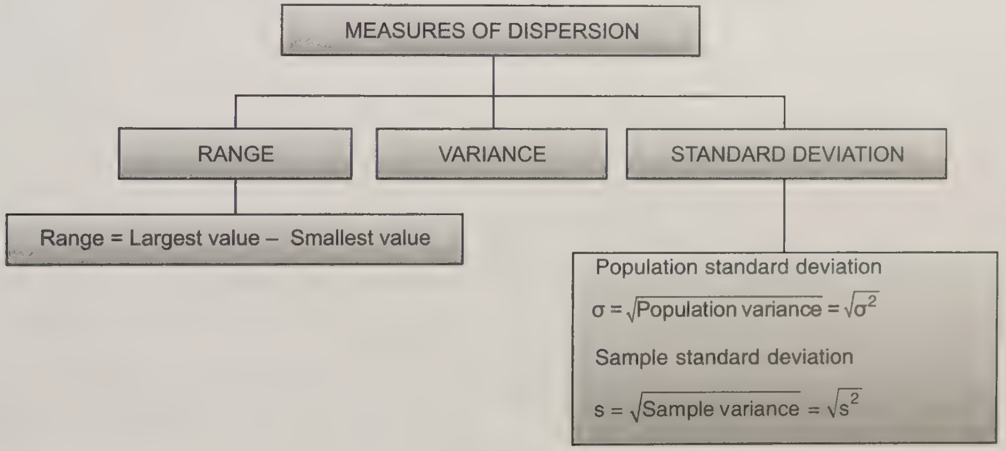

MEASURES OF DISPERSION
RANGE
VARIANCE
STANDARD DEVIATION
Range = Largest value – Smallest value
Population standard deviation
$\sigma = \sqrt{\text{Population variance}} = \sqrt{\sigma^2}$
Sample standard deviation
$s = \sqrt{\text{Sample variance}} = \sqrt{s^2}$

Fig. 1.5: Measures of dispersion

Example E1.8

The following data gives the number of years of service of 15 employees in a manufacturing company.

5, 9, 7, 6, 24, 11, 4, 13, 10, 9, 20, 8, 19, 17, 25

Calculate the range, variance and standard deviation.

SOLUTION:

Range = largest value – smallest value = 25 – 4 = 21 years

$$\text{Variance} = s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1} = \frac{2732 - \frac{178^2}{15}}{15-1} = \frac{2732 - 2112.266}{14} = 619.734$$

$$\text{Standard deviation} = s = \sqrt{s^2} = \sqrt{619.734} = 24.894$$

Example E1.9

The following data refers to the pay (in thousands of dollars) of 12 employees at a manufacturing company.

75	69	65	49	21	21	18	18	17	16	16	15

(*a*) calculate the mean, median and mode for this data

(*b*) find the range, variance and standard deviation.

SOLUTION:

(*a*) Mean $= \dfrac{\Sigma x}{n} = \dfrac{400}{12} = \$33{,}333$

Median is $\left[\dfrac{n+1}{2}\right]^{th}$ term in a ranked data set. Therefore,

$$\frac{n+1}{2} = \frac{12+1}{2} = 6.5$$

$$\text{Median} = \frac{21+18}{2} = \$19{,}500$$

Mode = \$21,000, \$18,000 and \$16,000

(*b*) Range = largest pay – smallest pay = \$75,000 – \$15,000 = \$60,000

$$\text{Variance} = s^2 = \frac{\Sigma x^2 - \dfrac{(\Sigma x)^2}{n}}{n-1} = \frac{19{,}528 - \dfrac{400^2}{12}}{12-1} = 563.1515$$

$$\text{Standard deviation} = s = \sqrt{s^2} = \sqrt{563.1515} = 23.73 = \$23{,}730$$

1.2.3 Mean, Variance and Standard Deviation for Grouped Data

1.2.3.1 Mean for Grouped Data

$$\text{Mean for population data: } \mu = \frac{\sum mf}{N} \tag{1.14}$$

$$\text{Mean for sample data: } \overline{X} = \frac{\sum mf}{n} \tag{1.15}$$

where, *m* is the midpoint and *f* is the frequency of a class.

1.2.3.2 Variance and Standard Deviation for Grouped Data

Population variance

$$\sigma^2 = \frac{\sum f\,(m-\mu)^2}{N} = \frac{\sum m^2 f - \dfrac{(\sum mf)^2}{N}}{N} \tag{1.16}$$

Sample variance

$$s^2 = \frac{\sum f\,(m-\overline{X})^2}{n-1} = \frac{\sum m^2 f - \frac{(\sum m f)^2}{n}}{n-1} \tag{1.17}$$

where σ^2 is the population variance, s^2 is the sample variance, m is the midpoint of a class, N is the population size, n is the sample size, μ is the population mean, $\overline{X}$ is the sample mean, and f the class frequency.

Population standard deviation $\sigma = \sqrt{\sigma^2}$ (1.18)

Sample standard deviation $s = \sqrt{s^2}$ (1.19)

Example E1.10

Calculate the mean, variance and standard derivation for the following population data.

x	0–3	4–7	8–11	12–15	16–19	20–23
f	7	4	18	12	7	6

SOLUTION:

Classes	f	m	mf	m^2 f
0-3	7	1.5	10.5	15.75
4-7	4	5.5	22	121
8-11	18	9.5	171	1624.5
12-15	12	13.5	162	2187
16-19	7	17.5	122.5	2143.75
20-23	6	21.5	129	2773.50
Total	54		617	8865.50

Population mean

$$\mu = \frac{\sum m f}{N} = \frac{617}{54} = 11.426$$

$$\text{Variance} = \sigma^2 = \frac{\sum m^2 f - \frac{(\sum m f)^2}{N}}{N} = \frac{8865.50 - \frac{617^2}{54}}{54} = \frac{8865.50 - 7049.796}{54} = 33.624$$

Standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{33.624} = 5.798$

1.2.4 Measures of Position

Here we examine several descriptive measures based on percentiles. Descriptive measures based on percentiles are not sensitive to the influence of a few extreme observations. Descriptive measures based on percentiles are generally preferred over those based on the mean and standard deviation.

A measure of position determines the position of a single value in relation to other values in a sample or a population data set. The measures of position considered here are the quartiles, percentiles, and percentile ranks as shown in Fig. 1.6.

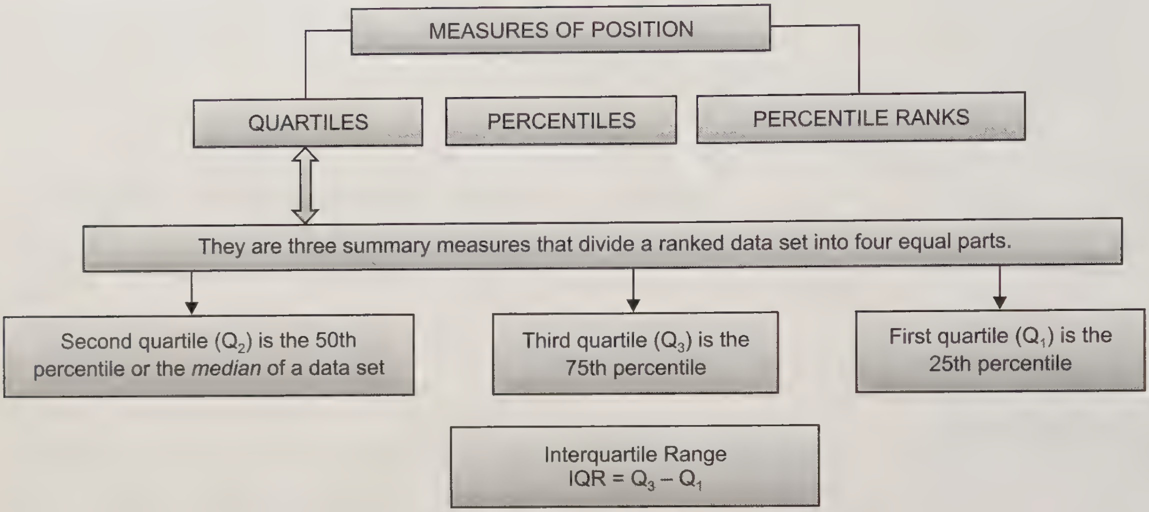

MEASURES OF POSITION
QUARTILES
PERCENTILES
PERCENTILE RANKS
They are three summary measures that divide a ranked data set into four equal parts.
Second quartile (Q_2) is the 50th percentile or the *median* of a data set
Third quartile (Q_3) is the 75th percentile
First quartile (Q_1) is the 25th percentile
Interquartile Range
IQR = $Q_3 - Q_1$

Fig. 1.6: Measures of position

1.2.4.1 Quartiles and Interquartile Range

The most commonly used percentiles are *quartiles*. Let n denotes the number of observations and if we arrange the data in an increasing order, then

- the first quartile is at position $\frac{n+1}{4}$
- the second quartile is the median, which is at position $\frac{n+1}{2}$
- the third quartile is at position $3\left(\frac{n+1}{4}\right)$

Thus, a data set has three quartiles, and denoted by Q_1, Q_2, and Q_3, Q_1, is the number that divides the bottom 25% of the data from the top 75%; the second quartile, Q_2 , the median, is the number that divides the bottom 50% of the data from the top 50%; the third quartile, Q_3, is the number that divides the bottom 75% of the data from the top 25%. Figure 1.7 shows the positions of the three-quartiles. Figure 1.8 shows the quartiles for uniform, bell-shaped, right-skewed, and left-skewed distributions. The difference between the third quartile and the first quartile of data set is called the *interquartile range* (IQR). Interquartile range is a preferred measure of variation when the median used as the measure of the center. The interquartile range is a resistant measure. IQR gives the range of the 50% of the observations.

Interquartile range IQR = $Q_3 - Q_1$

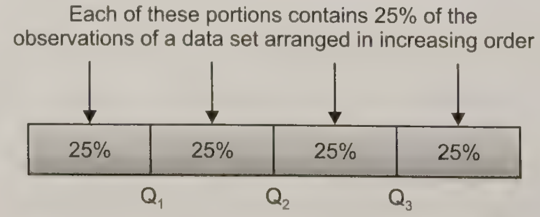
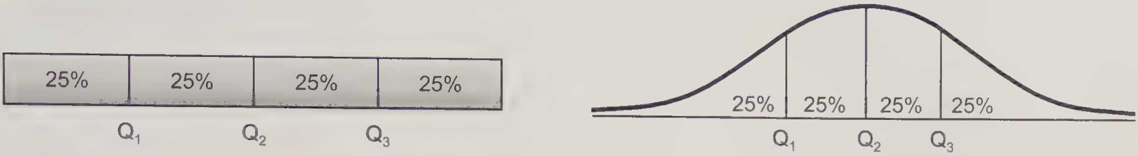
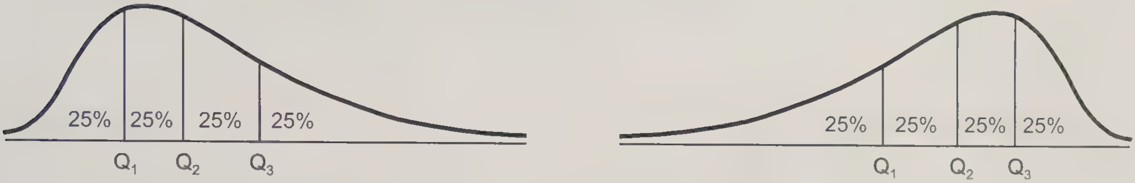

Each of these portions contains 25% of the observations of a data set arranged in increasing order
25%
25%
25%
25%
Q_1
Q_2
Q_3

Fig. 1.7: Quartiles



(a) Uniform

(b) Bell-shaped



(c) Right-skewed

(d) Left-skewed

Fig. 1.8: Quartiles for (a) uniform (b) bell-shaped (c) right-skewed and (d) left-skewed distributions

1.2.4.2 Percentiles

Percentiles are the summary measures that divide ranked data set into 100 equal parts. Each ranked data set has 99 percentiles that divide it into 100 equal parts. The k^{th} percentile is denoted by P_k, where k is an integer in the range 1 to 99. Figure 1.9 shows the positions of the 99 percentiles.

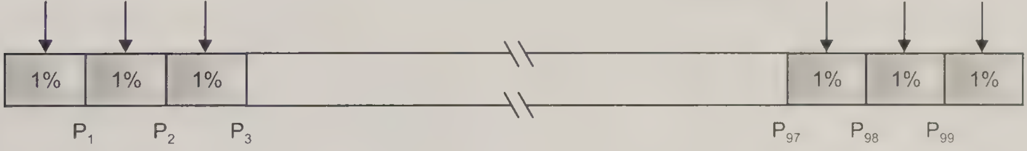

Each of these portions contains 1% of the observations of a data set arranged in increasing order

Fig. 1.9: Percentiles

Hence, the k^{th} percentile, P_k , can be described as a value in a data set such that about $k\%$ of the measurements are smaller than the value of P_k and about $(100 - k)\%$ of the measurements are greater than the value of P_k. The value of the k^{th} percentile is given by

$$P_k = \text{the value of the } \left(\frac{kn}{100}\right)^{th} \text{ term in a ranked data set}$$

where k denotes the number of the percentile and n represents the sample size.

The percentile rank for a particular value x_i of a data set is given by

$$\text{Percentile rank of } x_i = \frac{\text{number of values less than } x_i}{\text{total number of values in the data set}} \times 100$$

The percentile rank of x_i gives the percentage of values in the data set that are smaller than x_i.

Deciles: Deciles are also useful in statistics. The deciles of a data set divide into tenths, or 10 equal parts. A typical data set has nine deciles. If we denote them by D_1, D_2, ... D_9, then the first decile, D_1 is the number that divide the bottom 10% of the data from the top 90%; the second decile, D_2, is the number that divide the bottom 20% of the data from the top 80%; and so on. Hence, the first decile is the 10th percentile, the second decile is the 20th percentile, etc.

1.2.4.3 Skewness and Kurtosis

Skewness is the degree of asymmetry, or departure from symmetry, of a distribution. For skewed distribution, the mean tends to lie on the same side of the mode as the longer tail as shown earlier in Figs. 1.4(*a*) and (*b*).

Skewness is defined as

$$\text{Skewness} = \frac{\text{mean–mode}}{\text{standard deviation}} = \frac{\bar{X} - \text{mode}}{s} \tag{1.20}$$

or alternatively,

$$\text{Skewness} = \frac{3(\text{mean–median})}{\text{standard deviation}} = \frac{3(\bar{X} - \text{median})}{s} \tag{1.21}$$

The above two measures are called, respectively, *Pearson's first and second coefficient of skewness.*

Other definitions for skewness in terms of quartiles and percentiles are as follows:

$$\text{Quartile coefficient of skewness} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} \tag{1.22}$$

$$= \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1} \tag{1.22a}$$

$$\text{10–90 percentile coefficient of skewness} = \frac{(P_{90} - P_{50}) - (P_{50} - P_{10})}{P_{90} - P_{10}} \tag{1.23}$$

$$= \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}} \tag{1.23a}$$

If a frequency curve of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be *skewed to the right* or to have positive skewness. If the reverse is true, it is said to be *skewed to the left* or to have *negative skewness.*

Kurtosis is the degree of peakedness of a distribution taken relative to a normal distribution. A distribution having a relatively high peak such as the curve of Fig. 1.10(*a*) is called *leptokurtic*, while the curve of Fig. 1.10(*b*) which is flat-topped is called *platykurtic*. The normal distribution, Fig. 1.10(*c*), which is not very peaked or very flat-topped, is called *mesokurtic.*

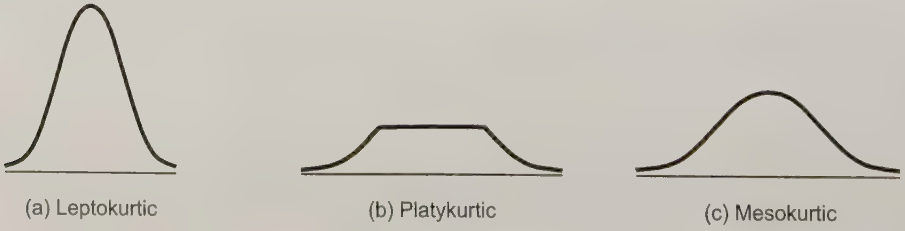

(a) Leptokurtic (b) Platykurtic (c) Mesokurtic

Fig. 1.10: Kurtosis

Example E1.11

The number of hours worked by 24 employees of a company is given below:

40	43	40	39	36	44	40	39
39	52	27	50	41	47	40	48
38	36	25	41	35	36	16	40

(a) calculate the three quartiles and the Interquartile range

(b) find the approximate value of the 81st percentile

(c) calculate the percentile rank of 37.

SOLUTION:

(a) The ranked data is

16 25 27 35 36 36 36 38 39 39 39 40 40 40 40 40 41 41 43 44 47 48 50 52

$$Q_2 = \text{median} = 40$$

$$Q_1 = \frac{36+36}{2} = 36$$

$$Q_3 = \frac{41+43}{2} = 42$$

IQR = Interquartile range = $Q_3 - Q_1 = 42 - 36 = 6$

(b) $\dfrac{k\,n}{100} = \dfrac{81(24)}{100} = 19.44$

Thus, the 81st percentile can be approximated by the value of the 19th term in the ranked data, which is 43. Hence $P_{81} = 45$.

(c) Seven values in the given data are smaller than 37.

Therefore, the percentile rank of 37 = $\dfrac{7}{24} \times 100 = 29.17\%$.

1.2.4.4 Box-and-Whisker Plot

A *box-and-whisker plot* depicts a graphical representation of data using five measures: the median, the first quartile, the third quartile and the smallest and the largest values in the data set between the lower and the upper inner fences. A box-and-whisker plot shows the center, spread, and skewness of a data set. It is constructed simply by drawing a box and two whiskers that use the median, the first quartile, the third quartile and the smallest and the largest values in the data set between the lower and the upper inner fences.

The box plot utilizes three fundamental measures of dispersion in a graphic manner for a set of data. The simplest measure of dispersion is the range. The box plot as shown in Fig.1.11 also incorporates the interquartile range, $Q_3 - Q_1$, and the semi-interquartile range, $(Q_3 - Q_1)/2$. The semi-interquartile range also corresponds to the median of the data set. Outliers are observations that fall well outside the overall pattern of the data. An outlier may be the result of a measurement or recording error, an observation from a different population, or an unusual extreme observations. We can utilize quartiles and the interquartile range (IQR) to identify potential outliers, that is, to spot observations that may be outliers. Referring to Fig.1.11 we define the lower and upper limits, the numbers that lie, respectively 1.5 IQRs below the first quartile and 1.5 IQRs above the third quartile.

Thus, Lower limit = $Q_1 - 1.5$ IQR

Upper limit = $Q_3 - 1.5$ IQR

Observations that lie outside the lower and upper limits are potential outliers. Typically outliers are observations that fall < ($Q_1 - 3$ IQR) and > ($Q_3 + 3$ IQR).

Box plots are useful for comparing two or more data sets and to identify the approximate shape of the distribution of a data set. Figure 1.12 shows the uniform, bell-shaped, right-skewed and left-skewed distributions and their corresponding box plots. The box width and whisker length relate to skewness and symmetry.

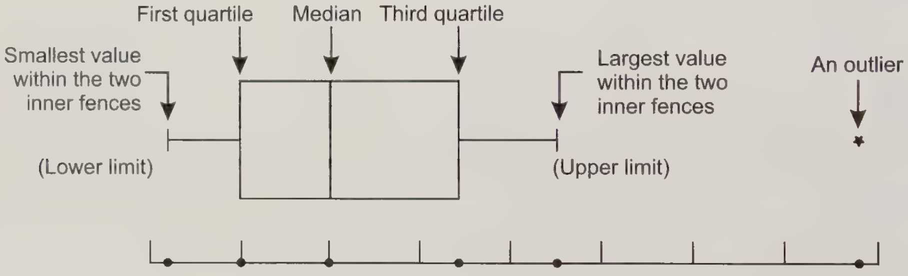
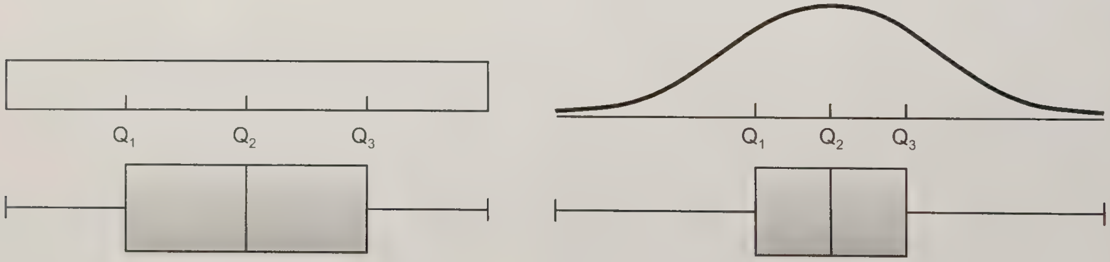
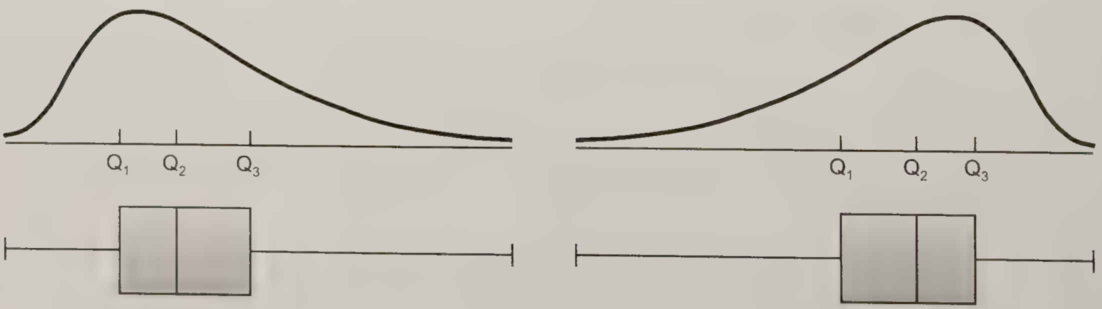

First quartile
Median
Third quartile
Smallest value within the two inner fences
(Lower limit)
Largest value within the two inner fences
(Upper limit)
An outlier

Fig. 1.11: Box-and-whisker plot



(a) Uniform

(b) Bell-shaped



(c) Right-skewed

(d) Left-skewed

Fig. 1.12: Distribution shapes and box plots for (a) uniform, (b) bell-shaped, (c) right-skewed, (d) left-skewed distributions

Example E1.12

Refer to the data in the Example E1.11 and construct a box-and-whisker plot and comment on the skewness of the data.

SOLUTION:

Median $= Q_2 = 40$, $Q_1 = 36$, $Q_3 = 42$, IQR $= Q_3 - Q_1 = 42 - 36 = 6$

Lower inner fence $= Q_1 - 9 = 36 - 9 = 27$

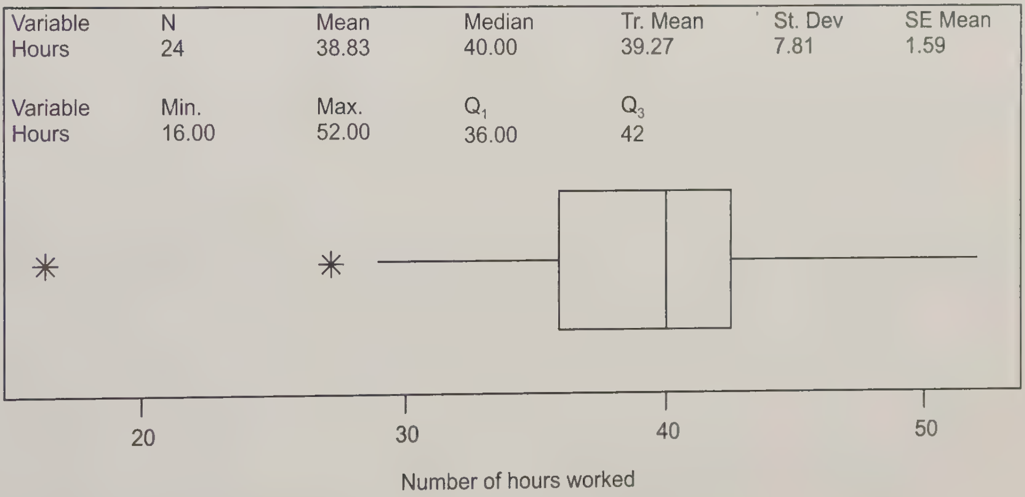
Upper inner fence $= Q_3 + 9 = 42 + 9 = 51$

The box-and-whisker plot is shown in Fig. E1.12.

The longest and smallest values with the two inner fences are 47 and 35 respectively. There are two outliers 50 and 52 shown by an asterisk in Fig. E1.12. The data in this example is skewed to the right, because the lower 50% of the values are spread over a smaller range than the upper 50% of the value.

Descriptive Statistics

Variable	N	Mean	Median	Tr. Mean	St. Dev	SE Mean
C1	24	38.83	40.00	39.27	7.81	1.59
Variable	Min.	Max.	Q_1	Q_3		
C1	16.00	52.00	36.00	42		



Variable	N	Mean	Median	Tr. Mean	St. Dev	SE Mean
Hours	24	38.83	40.00	39.27	7.81	1.59
Variable	Min.	Max.	Q_1	Q_3		
Hours	16.00	52.00	36.00	42		

Fig. E1.12: Box plot

1.3 SUMMARY

In this chapter, we have introduced the types of statistics, namely, the descriptive statistics and the inferential statistics. Descriptive statistics consists of methods for organizing and summarizing information. Inferential statistics consists of methods for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population.

PROBLEMS

P1.1 The following data give the results of a sample survey. The letters X, Y and Z represent the three categories.

X Y Y X Z Y Z Z Z X Z Y Z Z Z

Z Y X X X X Y X X Y Z Y Y Z X

(a) prepare a frequency distribution table

(b) calculate the relative frequencies and percentages for all categories

(c) what percentage of the elements in the sample belongs to category Y?

(d) what percentage of the elements in the sample belongs to category X and Z?

(e) draw bar graph for the frequency distribution.

P1.2 Fifteen faculty members at a university were asked whether ethics course should be required of all students as a requirement for graduation. The responses of these faculty members are listed below. (F, A, and N indicate that a faculty member is in favour, against, or has no opinion, respectively).

N A A F N N F F F A A N F F A

(a) prepare a frequency distribution table

(b) find the relative frequencies and percentages for all categories

(c) what percentage of the faculty members in this sample are in favour of this issue?

(d) draw a pie chart for the percentage distribution.

P1.3 The following data give the weights (in kg) of a random sample of 30 senior students at a university. Construct a frequency distribution table.

81 77 79 75 80 83 82 83 85 77 80 83 74 72 86

73 78 73 76 82 84 79 80 79 84 76 78 81 79 78

P1.4 A sample of randomly selected households in a large city produced the following data on the number of vehicles owned.

5 1 1 1 1 1 1 2 2 1 1 1 3 2 2

2 2 4 1 1 0 3 2 1 1 1 0 1 1 4

1 2 2 4 1 2 4 2 3

Draw a bar graph for vehicles owned.

P1.5 The following table gives the frequency distribution of ages of all 50 employees of a company.

Age	Number of employees
18 to 30	12
31 to 43	18
44 to 56	14
57 to 69	6

(*a*) find the class boundaries and class width

(*b*) do all classes have the same width?

(*c*) prepare the relative frequency and percentage distributions

(*d*) find the percentage of the employees that are 43 years old or younger.

P1.6 The following data give the number of machines manufactured by a company for a sample of 30 days.

24 21 31 32 26 35 27 31 34 23 22 22 33 27 26
33 33 31 29 27 23 25 23 35 23 28 31 28 29 27

(*a*) construct a frequency distribution table using the classes 21–23, 24–26, 27–29, 30–32 and 33–35.

(*b*) calculate the relative frequencies and percentages for all classes

(*c*) construct a histogram and a polygon for the percentage distribution.

(*d*) for what percentage of the days is the number of machines produced in the interval 27 to 29?

P1.7 The academic ranks of fifteen faculty members in a department are listed below. F, A and AS indicate that a faculty member is a Full Professor, Associate Professor and Assistant Professor, respectively.

F AS A F A A F AS F A F AS F A F

(*a*) prepare a frequency distribution table

(*b*) compute the relative frequencies and percentages for all categories

(*c*) what percentages of the faculty in this sample is full professor rank?

(*d*) draw a pie chart for the percentage distribution.

P1.8 The results of a sample survey of three categories 1, 2 and 3 are given below:

1 1 3 2 3 1 2 1 2 2 3 1 3
2 1 2 1 1 3 3 3 2 3 3 2 2

(*a*) prepare a frequency distribution table

(*b*) calculate the relative frequencies and percentages for all three categories

(*c*) find the percentage of the elements in this sample belonging to category 2

(*d*) find the percentage of the elements in this sample belonging to category 1 or 3.

P1.9 The following data gives the number of car oil changes made at a service station for a sample of 30 days.

23 31 26 24 34 34 30 26 24 27
21 27 30 23 28 32 27 23 27 30
29 35 33 23 27 27 24 35 30 28

(*a*) construct a frequency distribution table using the classes
21–23, 24–26, 27–29, 30–32 and 33–35

(*b*) calculate the relative frequencies and percentages for all classes

(*c*) construct a histogram and a polygon for the percentage distribution

(*d*) for what percentage of the days is the number of oil changes made in the interval 30–32?

P1.10 The following data gives the 1999 autocollision claims paid by an insurance company to 24 claimants.

2301	1678	4634	4541	851	2306	1174	974
1132	1828	749	2585	971	2891	1275	725
2782	1161	2406	912	3510	820	1560	1208

(a) construct the cumulative frequency, cumulative relative frequency and cumulative percentage distribution table

(b) draw an ogive for the cumulative percentage distribution

(c) using the ogive, find the percentage of claimants in this sample who received $2500 or less.

P1.11 The following data gives the collision repair cost in dollars for 25 automobiles randomly selected form a list of 100 cars.

2410	739	2500	413	556
2470	1789	2109	899	987
2170	326	1343	1158	1234
6181	4983	5892	2312	3953
6721	4800	1325	2911	1926

(a) construct a frequency distribution table. Assume the width of each class as $1500

(b) calculate the relative frequencies and percentages for all classes

(c) draw a histogram and a polygon for the relative frequency distribution

(d) what are the class boundaries and the width of the fourth class?

P1.12 The following data give the number of unemployed adults in 2008 for twenty states.

95	328	132	49	48	37	167	231	90	67
32	23	270	24	287	347	35	20	16	107

(a) construct a frequency distribution table. Take the classes as 1–70, 71–140 and so on

(b) prepare the relative frequency and percentages.

P1.13 The following data give the 2008 total annual compensation (in thousands of dollars) for 29 chief executive officers of some large corporations.

2595	1678	4832	4433	952	2207	1075	4200
1135	2829	767	2555	977	2891	1276	2781
1160	2406	908	3501	811	1568	964	727
1208	1029	1385	1011	2815			

(a) construct a frequency distribution table. Take $1 thousand as the lower limit of the first class and $1000 thousand as the width of the class

(b) prepare the relative frequency and percentages.

P1.14 The following data give the number of children less than 18 years of age for 30 randomly selected families.

2	1	1	1	2	0	2	0	0	2	1	2	3	0	1
1	2	2	1	1	3	2	2	2	2	0	0	0	0	0

(*a*) prepare a frequency distribution table using single-valued classes
(*b*) calculate the relative frequencies and percentages for all classes
(*c*) how many families in this sample have 2 or 3 children less than 18 years of age?
(*d*) draw a bar graph for the frequency distribution.

P1.15 Refer to Problem P1.5.
(*a*) prepare a cumulative frequency distribution table
(*b*) calculate the cumulative relative frequencies and cumulative percentages for all classes
(*c*) what percentage of the employees are 44 years of age or older?
(*d*) draw an ogive for the cumulative percentage distribution
(*e*) using the ogive, find the percentage of employees who are 40 years old or younger.

P1.16 Refer to Problem P1.6. Prepare the cumulative frequency, cumulative relative frequency and cumulative percentage distribution.

P1.17 Refer to Problem P1.13. Prepare the cumulative frequency, cumulative relative frequency and cumulative percentage distribution by using the table in Problem P1.13.

P1.18 Following are the SAT scores (out of a maximum possible score of 1600) of 12 students who took this test.

786	892	997	1170	881	1042
995	1084	773	980	1128	1066

Prepare a stem-and-leaf display. Arrange the leaves for each stem in increasing order.

P1.19 The following data give the time (in minutes) taken to commute from residence to work for 20 workers of accompany.

10	50	64	34	47	5	12	23	37	27
26	32	17	7	13	19	29	43	21	22

Construct a stem-and-leaf display for these data. Arrange the leaves for each stem in increasing order.

P1.20 At the end of a particular training program, 400 students in a college are judged by a panel of examiners as *excellent*, *very good*, *good*, or *fair*. The following data give the distribution:

Excellent	80
Very good	140
Good	120
Fair	60

Represent the figures by a pie chart.

P1.21 The following are the test scores of 60 students on engineering ethics test.

43	46	48	50	51	51	52	52	53	53
55	55	56	56	57	58	60	66	67	67
68	68	70	71	72	73	73	74	75	75
75	75	77	77	78	78	76	79	79	80
80	80	80	81	82	82	82	83	83	85
84	84	85	86	86	86	88	91	92	93

Construct a ranked stem-and-leaf display of the data.

P1.22 Plot the stem-and-leaf diagram for the following cholesterol readings (mg/deciliter) of twenty patients.

184 231 195 186 241 190 239 254 225 237
211 224 214 197 204 233 198 215 216 205

P1.23 The following are the scores of 30 students on math test:

75 53 80 96 66 78 72 87 93 95
69 72 81 62 76 86 79 68 50 92
83 84 77 64 71 87 72 93 58 99

Construct a ranked stem-and-leaf display of the data.

P1.24 The following data give the monthly rent paid by a sample of 30 households selected from a city.

428 586 733 675 550 989 1021 620 750 660
540 578 956 1030 1070 930 871 765 880 975
651 1020 950 840 781 870 900 801 750 821

Construct a stem-and-leaf display for the above data.

P1.25 The following data gives the moneys donated by 32 university faculty members to a charity during the calendar year 2008.

473 417 173 109 178 287 117 175
281 474 589 379 89 235 414 357
412 608 454 317 355 611 469 601
453 287 489 484 616 354 189 388

(a) prepare a stem-and-leaf display for these data using the last two digits as leaves

(b) condense the stem-and-leaf display by grouping the stem as 0–1, 2–3 and 4–6.

P1.26 The following data gives the number of hours spent exercising in a gymnasium by 10 randomly selected members of a university during the past week.

8 13 6 0 1 6 11 4 0 7

Compute the mean, median and mode.

P1.27 The following data gives the number of vacuum cleaners produced by a manufacturing company for a sample of 12 days.

25 33 28 24 36 34 30 22 24 29 31 23

Calculate the mean, median and mode.

P1.28 The following data gives the number of heavy machinery sold by a large manufacturing company for the 12 months in the year 2008.

3 2 47 118 204 97
68 86 62 57 98 99

Find the mean, median and mode.

P1.29 The following data are the test scores (in points) of 9 students in a class out 6 points:

5.8 5.7 5.9 5.7 5.5 5.7 5.7 5.7 5.6

Find the mean, median and mode for this data.

P1.30 The following data gives the number of ball bearings manufactured by a company per day for a sample of 10 days.

24 32 27 23 35 33 29 21 23 28

Calculate the mean, median and mode for these data.

P1.31 The following data give the number of automobile accidents in a city during the past 12 days.

3 7 11 5 6 3 8 7 6 9 13 2

Calculate the range, variance and standard deviation.

P1.32 Compute the range, variance and standard deviation for the data given in Problem P1.26.

P1.33 The following data give the number of hours spent doing the homework and attending classes by 10 randomly selected college students during the past week.

18 16 27 34 39 35 42 29 31 24

Compute the range, variance and standard deviation.

P1.34 The following data give the number of new automobiles sold at a dealership during a 15 days period.

13 5 9 7 6 10 11 14 12 8 4 8 16 10 8

Find the range, mean, variance and standard deviation.

P1.35 Following are the 2008 earnings (in thousands of dollars) before taxes of all 10 employees of a small company:

30 35 22 28 19 47 52 43 39 32

Calculate the variance and standard deviation of these data.

P1.36 The following data give the number of automobiles that stopped at a service station during each of the 10 hours observed:

30 32 35 29 28 17 19 30 18 25

Find the range, variance and standard deviation.

P1.37 Consider the following two data sets:

Data set 1:	8	12	25	37	41
Data set 2:	15	19	32	44	48

Calculate the standard deviation of each data set. Comment on the results obtained.

P1.38 The following table gives the frequency distribution of the amounts of long distance telephone bills for the month of August 2008 for a sample of 60 families in a city.

Amount of long distance telephone bill (dollars)	Number of families
25 to less than 50	15
50 to less than 75	12
75 to less than 100	20
100 to less than 125	8
125 to less than 150	5

Calculate the mean, variance and standard deviation.

P1.39 Using the sample formulae, find the mean, variance and standard deviation for the grouped data given in the following table:

x	f
0 to less than 4	16
4 to less than 8	23
8 to less than 12	14
12 to less than 16	10
16 to less than 20	8
20 to less than 24	7

P1.40 The following table gives the frequency distribution of entertainment expenditure (in dollars) incurred by 50 families during the past week.

Entertainment expenditure (dollars)	Number of families
0 to less than 10	5
10 to less than 20	10
20 to less than 30	14
30 to less than 40	10
40 to less than 50	6
50 to less than 60	5

Find the mean, variance and standard deviation.

P1.41 The following table gives the grouped data on the weights of all 100 babies born at a hospital in 2005.

Weight (kg)	Number of babies
1 to less than 2	5
2 to less than 3	30
3 to less than 4	40
4 to less than 5	20
5 to less than 6	5

Find the mean, variance and standard deviation.

P1.42 The following table gives the frequency distribution of the number of personal computers sold during the past month at 40 computer stores in a big city. Calculate the mean, variance and standard deviation.

Computers sold	Number of stores
4 to 12	6
13 to 21	10
22 to 30	13
31 to 39	6
40 to 48	5

P1.43 The following table gives the number of television sets owned by 80 households.

Number of television sets owned	Number of households
0	4
1	32
2	27
3	11
4	6

Find the mean, variance, and standard deviation. Use the number of television sets owned as the values of m in the formulae.

P1.44 The following data gives the scores of 12 students in Design for Economy and Reliability class.

74 79 67 52 98 57 75 72 84 87 90 78

(a) calculate the values of the three quartiles and the interquartile range. Where does the value of 80 lies relative to these quartiles?

(b) calculate the value of the 93rd percentile. Comment on the calculated percentile.

(c) determine the percentile rank of 81. Comment on the percentile rank.

P1.45 The following data gives the number of hours worked in the last week by 30 employees of a manufacturing company.

38 40 42 45 35 47 28 40 43 39

53 40 51 23 42 49 40 51 36 40

49 33 21 31 40 33 39 16 36 41

(a) calculate the values of the three quartiles and the interquartile range

(b) calculate the approximate value of the 79th percentile

(c) calculate the percentile rank of 39.

P1.46 The following data gives the number of electronic components produced at ABC company for a sample of 30 days.

24 32 27 23 33 33 29 23 25 28

20 20 26 31 33 27 27 28 29 23

36 31 26 34 22 36 23 28 31 27

(a) calculate the values of the three quartiles and the interquartile range. Where does the value of 31 lie in relation to these quartiles?

(b) calculate the value of the 65th percentile

(c) calculate the percentage of days when the number of electronic components produced was 32 or higher.

P1.47 The following data gives the number of new computer systems sold during a 19 days period.

93 84 75 63 97 81 73 69 47 58
95 78 84 77 45 69 83 98 84

(a) calculate the values of the three quartiles and the interquartile range. Where does the value of 81 lies in relation to these quartiles?
(b) calculate the approximate value of the 93rd quartile
(c) find the percentile rank of 82.

P1.48 The following are the scores of 20 students in a Mathematics class:

48 82 82 83 70 73 78 74 46 81
78 74 64 80 79 80 80 79 80 80

(a) find the quartiles
(b) find the interquartile range
(c) find the five number summary.

P1.49 The random sample of 21 tourists yielded the following data on length of stay (in days) in a summer resort.

4 3 10 4 6 13 12 15 5 18 7
7 12 2 9 6 25 23 1 3 9

(a) find the quartiles
(b) find the interquartile range
(c) find the five number summary.

P1.50 The following are the scores of 12 students in a statistics class.

74 81 68 51 97 57 76 73 85 88 90 79

(a) find the values of the 3 quartiles
(b) find the interquartile range.

P1.51 The following are the ages of 9 employees of a small manufacturing company.

47 28 40 51 34 37 62 23 33

(a) find the values of the three quartiles
(b) find the interquartile range.

P1.52 Refer to Problem P1.50. Find the value of the 62nd percentile.

P1.53 Refer to Problem P1.50. Find the percentile rank for the score 85.

P1.54 A sample of 20 full-time students in a college yielded their weekly times spent on attending classes and doing their homework in hours. Determine the quartiles for these data.

43 6 16 21 32 14 38 20 27 31
32 30 25 64 34 30 41 35 26 38

P1.55 Refer to Problem P1.54.

(a) find the five-number summary
(b) find the interquartile range
(c) obtain the lower and upper limits
(d) identify potential outliers, if any.

P1.56 The following data gives the incomes in thousands of dollars for a sample of 12 employees in a company

22 16 31 59 21 51 28 37 41 91 26 45

Construct a box-and-whisker plot for this data set.

P1.57 The following table shows the number of used cars sold each week during 10 weeks period by a car dealership in a large city.

66 96 147 147 154 175 88 154 57 116

(a) find the 3 quartiles

(b) determine the interquartile range

(c) identify the potential outliers, if any

(d) construct a box plot.

P1.58 The following data give the time (in minutes) taken by 14 students to complete a Maths test.

73 91 82 71 74 98 95 89 88 79 93 96 87 77

Prepare a box-and-whisker plot. Comment on the skewness of these data.

P1.59 The following data give the income (in thousands of dollars) for a sample of 12 households.

17 23 32 22 52 60 29 38 92 42 27 46

Construct a box-and-whisker plot for these data.

P1.60 The following data give the hours worked by 30 employees of a company during one week.

43 39 27 40 42 40 35 45 38 47

40 51 36 40 40 23 42 53 51 48

36 41 39 16 48 21 31 34 40 34

Make a box-and-whisker plot.

P1.61 A corporation manufactures ball bearings. The following data give the number of bearings produced at the company for a sample of 30 days.

31 35 22 34 23 28 26 27 31 35

24 27 32 23 33 33 29 23 28 25

21 26 20 31 33 27 23 27 29 28

Prepare a box-and-whisker plot. Comment on the skewness of these data.

REVIEW QUESTIONS

1. Describe the need to group data in the form of a frequency table.
2. Describe how the relative frequencies and percentages of categories obtained from the frequencies of categories.
3. Describe the steps to be made to group a data set in the form of a frequency distribution table.
4. How are the relative frequencies and percentages of classes obtained from the frequencies of classes?

5. Describe the concept of cumulative frequency distribution. How are cumulative relative frequencies and cumulative percentages calculated?
6. Briefly explain how a stem-and-leaf display for a data set is prepared.
7. Describe very briefly the following:
    - *(a)* bar graph
    - *(b)* class
    - *(c)* class boundary
    - *(d)* class frequency
    - *(e)* class width or size
    - *(f)* grouped data
    - *(g)* histogram
    - *(h)* ogive
    - *(i)* pie chart
8. Explain the difference between ungrouped and grouped data.
9. Briefly explain the meaning of the following terms:
    - *(a)* median
    - *(b)* outlier
    - *(c)* mean
    - *(d)* mode
10. Explain the relationship between mean, median and mode for symmetric and skewed histograms.
11. Explain the difference between a population parameter and a sample statistic.
12. Are the values of the mean and standard deviation that are calculated using grouped data exact or approximate values of the mean and standard deviation, respectively? Explain.
13. Describe how the three quartiles are calculated for a data set.
14. What is interquartile range?
15. Describe how the percentiles are calculated for a data set.
16. Describe the concept of the percentile rank for an observation of a data set.
17. Explain what summary measures are used to construct a box-and-whisker plot.
18. Define the following terms:
    - *(a)* coefficient of variation
    - *(b)* measures of dispersion
    - *(c)* measures of position
    - *(d)* parameter
    - *(e)* range
    - *(f)* first quartile
    - *(g)* second quartile
    - *(h)* statistic
    - *(i)* third quartile
    - *(j)* upper inner fence
    - *(k)* upper outer fence
    - *(l)* variance

STATE TRUE OR FALSE

1. Variable is a characteristic that varies from one person or thing to another. (True/False)
2. Qualitative variable is a numerically valued variable. (True/False)
3. Quantitative variable is a numerically valued variable. (True/False)
4. Discrete variable is a qualitative variable whose possible values can be listed. (True/False)
5. Continuous variable is a quantitative variable whose possible values form some interval of numbers. (True/False)
6. Data is values of a variable. (True/False)
7. Qualitative data is values of a qualitative variable. (True/False)

8. Continuous data is values of a continuous variable. (True/False)
9. Qualitative variable yields non-numerical data. (True/False)
10. Classes are categories for non-grouping data. (True/False)
11. Frequency is the number of observations that fall in a class. (True/False)
12. Frequency distribution is a listing of all classes and their relative frequencies. (True/False)
13. Relative-frequency distribution is a listing of all classes and their frequencies. (True/False)
14. Lower cut point is the smallest value that could go in a class. (True/False)
15. Upper cut point is the smallest value that could go in the next higher class (equivalent to the lower cut point of the next higher class). (True/False)
16. The middle of a class is found by summing its cut points. (True/False)
17. Width is the average between the cut points of a class. (True/False)
18. Grouping can help to make a large and complicated set of data more compact and easier to understand. (True/False)
19. The concepts of cut points and mid points make sense only for numerical data (for which doing arithmetic is meaningful). (True/False)
20. The frequency of a class is the number of observations in the class, whereas the relative frequency of a class is the ratio of the class frequency to the total number of observations. (True/False)
21. The percentages of a class equals 100 times the relative frequency of the class. (True/False)
22. The relative frequency of a class is the percentage of the class expressed as a decimal.(True/False)
23. Two data sets that have identical frequency distributions have identical relative-frequency distributions. (True/False)
24. Two data sets that have identical relative-frequency distributions have identical frequency distributions. (True/False)
25. The four elements of a grouped-data table are the classes, frequencies, relative frequencies, and mid points. (True/False)
26. The mid point of each class is not the same as the class. (True/False)
27. Relative frequencies always lie between 0 and 1 and hence provide a standard for comparison. (True/False)
28. A frequency histogram displays the class frequencies on the vertical axis, whereas a relative-frequency histogram displays the class relative frequencies on the vertical axis. (True/False)
29. Stem-and-leaf diagrams are generally useful with large data sets. (True/False)
30. The distribution of a data set is a table, graph, or formula that provides the values of the observations and how often they occur. (True/False)
31. Sample data are the values of a variable for a sample of the population. (True/False)
32. Population data are the values of a variable for an entire population. (True/False)
33. Census data is another name for sample data. (True/False)
34. A sample distribution is the distribution of population data. (True/False)

35. The population distribution is the distribution of population data. (True/False)
36. Distribution of a variable is another name for the population distribution. (True/False)
37. The measure of center is to indicate where the center or most typical value of data set lies. (True/False)
38. The three most important measures of center are the mean, the median and the mode. (True/False)
39. The mean, median and mode can all be used with quantitative data. (True/False)
40. If the number of observations is even, then the median is the observation exactly the middle of the ordered list. (True/False)
41. If the number of observations is odd, then the median is the mean of the two middle observations in the ordered list. (True/False)
42. The range of a data set is the difference between the maximum (largest) and minimum (smallest) observations. (True/False)
43. The more variation that there is in a data set, the larger is its standard deviation. (True/False)
44. Almost all the observations in any data set lie within three standard deviations to either side of the mean. (True/False)
45. The purpose of a measure of variation is to indicate the amount of variation in a data set. (True/False)
46. The first quartile is the median part of the entire data set that lies at or below the median of the entire data set. (True/False)
47. The second quartile is the median of the entire data set. (True/False)
48. The third quartile is the median part of the entire data set that lies at or below the median of the entire data set. (True/False)
49. The interquartile range is the difference between the first and third quartiles. (True/False)
50. The median and interquartile range are resistant measures, whereas the mean and standard deviations are not. (True/False)
51. Box plots are useful when the data set (number of observations) is large. (True/False)
52. Parameter is a descriptive measure for a population. (True/False)
53. Statistic is a descriptive measure for a sample. (True/False)
54. For an observed value of a variable x, the corresponding value of the standardised variable z is called the x-score of the observation. (True/False)
55. Numbers that are used to describe data sets are called descriptive measures. (True/False)
56. Descriptive measures that indicate where the center or most typical value of a data set lies are called measures of dispersion. (True/False)
57. Descriptive measures that indicate the amount of variation or spread, in a data set are called measures of variation. (True/False)
58. Almost all the observations in any data set lie within 3 standard deviations to either side of the mean. (True/False)
59. An outlier is an observation that falls well outside the overall pattern of the data. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **False** 5. **True** 6. **True** 7. **True** 8. **True** 9. **True** 10. **False**

11. **True** 12. **False** 13. **False** 14. **True** 15. **True** 16. **False** 17. **False** 18. **True** 19. **True** 20. **True**

21. **True** 22. **True** 23. **True** 24. **True** 25. **True** 26. **False** 27. **True** 28. **True** 29. **False** 30. **True**

31. **True** 32. **True** 33. **False** 34. **False** 35. **True** 36. **True** 37. **True** 38. **True** 39. **True** 40. **False**

41. **False** 42. **True** 43. **True** 44. **True** 45. **True** 46. **True** 47. **True** 48. **False** 49. **True** 50. **True**

51. **True** 52. **True** 53. **True** 54. **False** 55. **True** 56. **False** 57. **True** 58. **True** 59. **True**

CHAPTER 2

Probability

In Chapter-1, we have discussed on the descriptive statistics, that is, methods for organising and summarising numerical data. Another important aspect is to present the fundamentals of inferential statistics, that is, methods of drawing conclusions about a population based on information from a sample of the population. Because inferential statistics involves utilising information from part of a population (a sample) to infer conclusions about the entire population, we can never be certain that our conclusions (inferences) are correct or true, that is, uncertainty is inherent in inferential statistics. Therefore, we need to be familiar with uncertainty before we can understand, develop, and apply the methods of inferential statistics.

The science of uncertainty is called *probability theory*. Probability theory enables us to evaluate and control the likelihood that a statistical inference is correct. More generally, probability theory provides the mathematical basis for inferential statistics.

This chapter introduces the basic concepts and definitions on probability, events (simple and compound), Venn diagram, tree diagram, approaches to probability (classical, relative frequency concept of probability, subjective probability, marginal probability and conditional probability. Special multiplication rule and Baye's formula are also briefly presented.

2.1 EXPERIMENT, OUTCOME AND SAMPLE SPACE

The tossing of a coin or the rolling of a die constitutes an *experiment*. In probability and statistics, the term *experiment* is used in a very wide sense and refers to any procedure that yields a collection of outcomes. The knowledge of all possible outcomes when a coin is tossed, or a die is rolled, is important. This is always the case for determining the probabilities. Measuring the length of a bolt, weighing the contents of a box of materials, and measuring the breaking strength of a metal component are all examples of experiments.

An *experiment* is a process that, when performed, results in one and only one of many observations. These observations are known as the *outcomes* of the experiment. The collection of all outcomes for an experiment is called a *sample space*. For tossing a coin, we can use the set {Heads, Tails} as the sample space. For rolling a six-sided die, we can use the set {1, 2, 3, 4, 5, 6}. These sample spaces are finite. Some experiments have sample spaces with an infinite number of outcomes. The elements of a sample space are called *sample points*. The *sample space* for an experiment can be described by drawing either a *Venn diagram* or a *tree diagram*. A *Venn diagram* is a picture that depicts all the possible outcomes for an experiment. In a *tree diagram*, a branch of the tree represents each outcome.

2.2 SIMPLE AND COMPOSITE EVENTS

An *event* is a collection of one or more of the outcomes of an experiment. An event may be simple event or a compound event as shown in Fig. 2.1. A simple event is also called an *elementary event*, and a composite event is also called a *compound event*. An event that includes one and only one of the final outcomes for an experiment is called a *simple event* and is generally denoted by E_i. A compound event is a collection of more than one outcome of an experiment. The probability of a composite event is the sum of the probabilities of the simple events of which it is composed. When all the sample points or simple events of an experiment are equally likely and so have equal probabilities, then the probability of a composite event is easily found.

For instance, if there are k sample points and a composite event A contains r of them, then $P(A) = \frac{r}{k}$.

The probability that a simple event E_1 will occur is denoted by $P(E_1)$, and the probability that a compound event A will occur is denoted by $P(A)$.

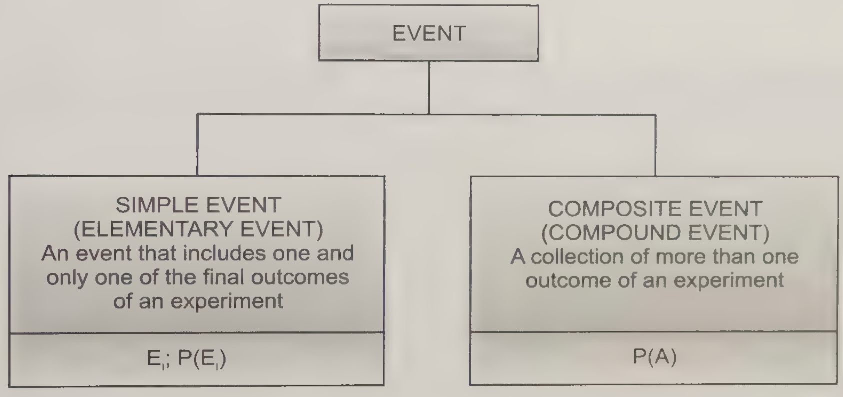

EVENT
SIMPLE EVENT
(ELEMENTARY EVENT)
An event that includes one and only one of the final outcomes of an experiment
E_i; P(E_i)
COMPOSITE EVENT
(COMPOUND EVENT)
A collection of more than one outcome of an experiment
P(A)

Fig. 2.1: Classification of an event

An *impossible event* is one that has no outcomes in it and consequently, cannot occur. On the other hand, a *sure event* is one that has all the outcomes of the sample space in it and will, therefore, definitely occur when the experiment is performed. Thus, the sample space constitutes a sure event.

Probability Definition: Let us assume that the sample space S has N outcomes e_1, e_2, e_N so that there are N simple events $\{e_1\}, \{e_2\}, \{e_N\}$.

The *probability of a simple event* $\{e\}$ is a number denoted by $P[\{e\}]$ and satisfies the following conditions:

1. $P[\{e\}]$ is always between 0 and 1 that is $0 \leq P[\{e\}] \leq 1$.
2. The sum of the probabilities of all the simple events is 1; that is, $P[\{e_1\}] + P[\{e_2\}] + \cdots + P[\{e_N\}] = 1$

The *probability of an event* A, denoted by $P(A)$, is defined as the sum of the probabilities assigned to the simple events that comprise the event A. The impossible event has probability 0 and the sure event has probability 1.

Example E2.1

Find the probability of getting *(a)* exactly two tails (event A), *(b)* at least two tails (event B) in tossing 3 balanced fair coins.

Solution:

(a) The event A of exactly two tails comprises of the sample points (*TTH, THT, HTT*).

Hence, $$P(A) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}$$

(b) The event B of at least two tails comprises of the sample points (*TTT*), (*TTH*), (*THT*) and (*HTT*).

Hence, $$P(B) = \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}$$

Example E2.2

Find the probability that the sum of the numbers shown in the two faces, when two dice are thrown *(a)* is 8 and *(b)* 10.

Solution:

(a) The event A that the sum of the numbers shown on the two faces is eight consists of the sample points:

(1, 7), (2, 6), (3, 5), (4, 4), (5, 3), (6, 2) and (7, 1)

There are 7 sample points. Therefore, the required probability is $\frac{7}{36}$.

(b) The event that the sum is 10 consists of the following sample points:

(4, 6), (5, 5) and (6, 4)

Hence, the required probability is $\frac{3}{36} = \frac{1}{12}$.

2.3 AXIOMS OF PROBABILITY

The subject of probability is based on three rules, known as axioms. They are:

1. Let S be a sample space. Then $P(S) = 1$.
2. For any event A, $0 \le P(A) \le 1$.
3. If A and B are mutually exclusive events, then $P(A \cup B) = P(A) + P(B)$

   More generally, if $A_1, A_2, A_3, \ldots$ are mutually exclusive events then $P(A_1 \cup A_2 \cup A_3 \cup \cdots) = P(A_1) + P(A_2) + \cdots$

2.4 FINITE PROBABILITY SPACES

Consider a sample space S and the class C of all events. We assume here that if S is finite, then all subsets of S are events. S then become a probability space by assigning probabilities to the events in C so that they satisfy the probability axioms.

If S is a finite sample space with n elements and suppose the physical characteristics of the experiment suggest that the various outcomes of the experiment be assigned equal probabilities. Then S becomes a probability space, called a *finite equiprobable space*, if each point in S is assigned the probability $1/n$ and if each event A containing r points is assigned the probability r/n.

$$\text{Hence,}\quad P(A) = \frac{\text{number of elemens in } A}{\text{number of elements in } S} = \frac{n(A)}{n(S)} \tag{2.1}$$

$$\text{or}\qquad P(A) = \frac{\text{number of ways that the event } A \text{ can occur}}{\text{number of ways that the sample sapce } S \text{ can occur}} \tag{2.2}$$

Let $S = \{a_1, a_2, \ldots, a_n\}$ be a finite sample space. Then, a *finite probability space* or *finite probability model* is obtained by assigning each point a_i in S a real number p_i, called the probability of a_i, satisfying the following properties:

1. Each p_i is nonnegative, $p_i \geq 0$.
2. The sum of the $p_i = 1$, that is, $p_1 + p_2 + \cdots + p_n = 1$.

The *probability* $P(A)$ of an event A is defined to be the sum of the probabilities of the points in A.

2.5 INFINITE PROBABILITY SPACES

There are two cases: S is countably infinite and S is uncountable. A finite or countably infinite probability space S is said to be *discrete*. An uncountable space S which consists of a continuum of points is said to be *continuous*.

Case 1: Countably Infinite Sample Spaces

Suppose S is a countably infinite sample space: $S = \{a_1, a_2, \ldots\}$. Then, a probability space is obtained by assigning $a_i \in S$ a real number p_i, called its *probability*, such that

1. Each p_i is nonnegative, or $p_i \geq 0$.
2. The sum of the p_i is equal to 1

$$\text{or } p_1 + p_2 + \cdots = \sum_{i=1}^{\infty} p_i = 1.$$

The probability $P(A)$ of an event A is then equals the sum of the probabilities of its points.

Case 2: Uncountable Spaces

The probability of an event A is given by the ratio of $m(A)$ to $m(S)$:

$$P(A) = \frac{\text{length of } A}{\text{length of } S}$$

$$\text{or}\qquad P(A) = \frac{\text{area of } A}{\text{area of } S} \tag{2.3}$$

$$\text{or}\qquad P(A) = \frac{\text{volume of } A}{\text{volume of } S}$$

where the uncountable sample spaces S are those with some finite geometrical measurement such as length, area, or volume and in which a point is selected at random. Such a probability space is said to be *uniform*.

2.6 PROPERTIES OF PROBABILITY

The basic properties of probabilities are:

1. The probability of an event always lies in the range of 0 and 1. That is

   $0 \leq P(E_1) \leq 1$ Simple event

   $0 \leq P(A) \leq 1$ Compound event

2. The sum of the probabilities of all simple events or final outcomes for an experiment, denoted by $\Sigma P(E_i)$, is always equal to 1.

   Hence,
   $$\Sigma P(E_i) = P(E_1) + P(E_2) + \cdots = 1 \tag{2.4}$$

3. The probability of an event that cannot occur is 0. (An event that cannot occur is called an *impossible event*).
4. The probability of an event that must occur is 1. (An event that must occur is called a *certain event*).

2.7 VENN DIAGRAM

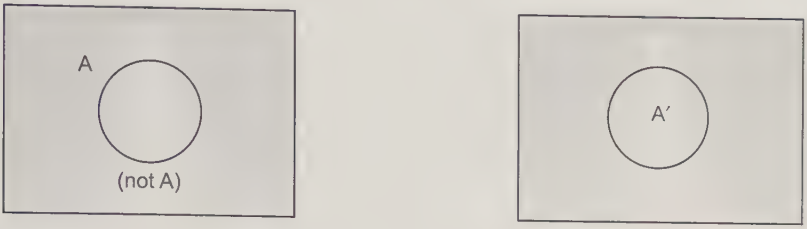
Graphical display of events are helpful for explaining and understanding probability. Venn diagrams, named after English logician John Venn (1834–1923) are considered one of the excellent ways to visually display events and relationships among events. The sample space is shown as a rectangle and the various events are drawn as circles (or other geometric shapes) inside the rectangle. Venn diagram for one event is shown in Fig. 2.2.

Each event A has a corresponding event defined by the condition that "A does not occur". That event is called the *complement* of A and is denoted (not A) or (A') as shown in Fig. 2.2(b). Event (not A) consists of all outcomes not in A, as shown in the Venn diagram in Fig. 2.2(a). With any two events, say, A and B, we can associate two new events. One new event is defined by the condition that "both event A and event B occur" and is denoted (A & B) or (A and B). Event (A and B) consists of all outcomes common to both event A and event B, as shown in Fig. 2.2(d). The other new event associated with A and B is defined by the condition "either event A or event B or both occur" or equivalently, that, "at least one of events A and B occur". That event is denoted (A or B) and consists of all outcomes in either event A or event B or both, as shown in Fig. 2.2(c).

Relationship among Events

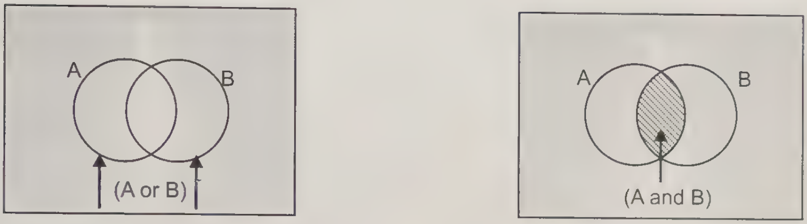
(not A):	The event "A does not occur".
(A and B) or (A & B):	The event "both A and B occur".
(A or B):	The event "either A or B or both occur".

Note here that event "both A and B occurs" is the same as the event "both B and A occur", event (A and B) or (A & B) is the same as event (B and A) or (B & A). Similarly, event (A or B) is the same as event (B or A).



(a) Event A

(b) Complement of event A(A′)



(c) Union of events A and B (A∪B)
Event A or B

(d) Intersection of events A and B (A∩B)
Event (A and B)

Fig. 2.2: Venn diagram

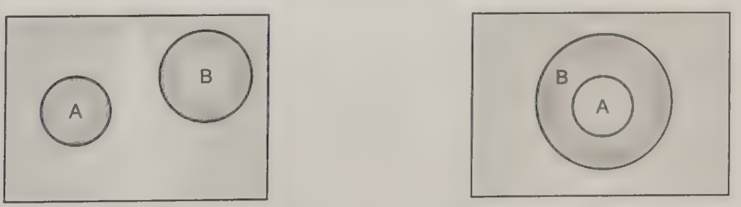
2.8 PROBABILITY TREE OR TREE DIAGRAM

In a *(finite) stochastic process*, in which a finite sequence of experiments where each experiment has a finite number of outcomes with given probabilities, the most convenient way to describe such a process is by means of a *labeled tree diagram.*

Various events are drawn as circles (or other geometric shape such as a rectangle, a square, or a circle) that depicts all the possible outcomes for an experiment. For instance, the shaded regions of the four Venn diagrams of Fig. 2.2 represents respectively, event A, the complement of event A, the union of events A and B, and the intersection of events A and B.

If A and B are any two subsets of a sample space S, their union $A \cup B$ is the subset of S that contains all the elements that are either in A, in B, or in both; their intersection $A \cap B$ is the subset. S that contains all the elements that are in both A and B; on the complement A' of A is the subset of S that contains all the elements of S that are not in A.

Figure 2.3(*a*) indicates that events A and B are *mutually exclusive*, that is, the two sets have no elements in common (or the two events cannot both occur). This is written $A \cap B = \phi$ denotes the *empty set*, which has no elements at all. Figure 2.3(*b*) shows that A is contained in B, we write this as $A \subset B$.



(a) Events A and B mutually exclusive

(b) Event A contains in Event B

Fig. 2.3: Venn diagram sharing special relationships among events A and B

Example E2.3

A bin contains a certain number of manufactured mechanical components, a few of which are defective. Two components are selected at random from this bin and inspected to determine, if they are non-defective or defective. How many total outcomes are possible? Draw a tree diagram for this experiment. Show all the outcomes in a Venn diagram.

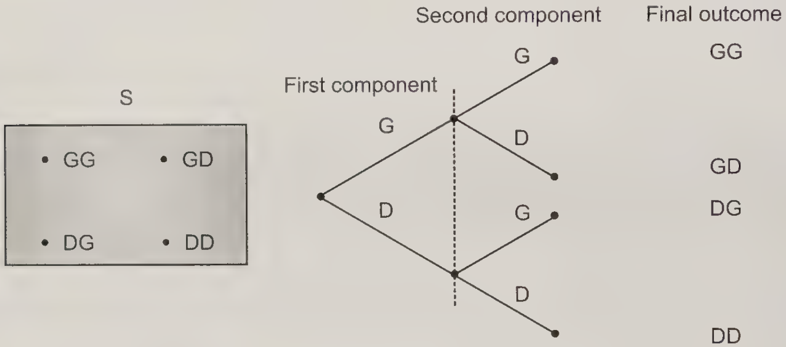
SOLUTION:

Let G = the selected component is good

D = the selected component is defective

The four outcomes for this experiment are: GG, GD, DG and DD.

The Venn and tree diagrams are shown in Figs. E2.3(*a*) and (*b*).



(a) Venn diagram (b) Tree diagram

Fig. E2.3: Venn and tree diagrams

Example E2.4

In Fig. E2.4, H is the event that an employee has health insurance and D is the event that the employee has disability insurance.

(*a*) express in words what events are represented by regions 1, 2, 3 and 4

(*b*) what events are represented by regions 1 and 2 together

(*c*) regions 2 and 4 together

(*d*) regions 1, 2 and 3 together

(*e*) regions 2, 3 and 4 together.

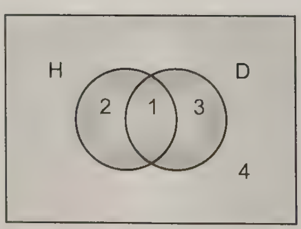


Fig. E2.4: Venn diagram

SOLUTION:

(a) 1 : The employee has health insurance and disability insurance

2 : The employee has health insurance but no disability insurance

3 : The employee has disability insurance but no health insurance

4 : The employee has neither health insurance nor disability insurance

(b) The employee has health insurance.

(c) The employee does not have health insurance.

(d) The employee has either health or disability insurance, but not both.

(e) The employee does not have both kinds of insurance.

Example E2.5

A small box contains a few red and a few blue balls. If two balls are randomly drawn and the colours of these balls are observed, how many total outcomes are possible? Draw a tree diagram for this experiment. Show all the outcomes in a Venn diagram.

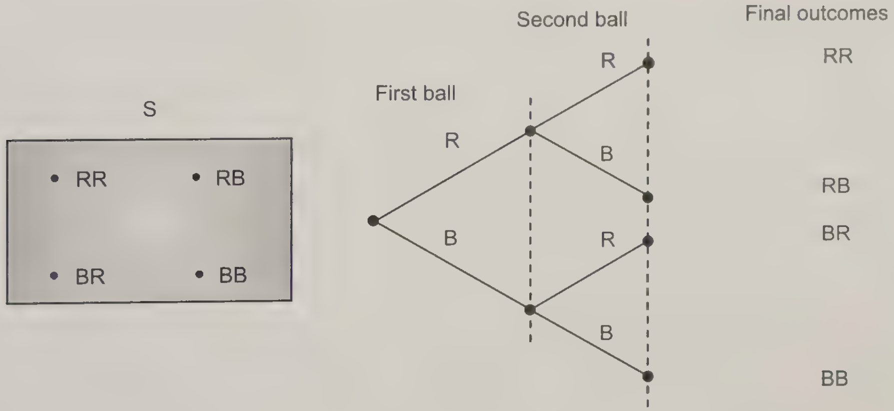
SOLUTION:

Let R = a red ball is selected

B = a blue ball is selected

The experiment has 4 outcomes: RR, RB, BR and BB.

The Venn diagram and tree diagram are shown in Figs. E2.5(*a*) and (*b*).



(a) Venn diagram

(b) Tree diagram

Fig. E2.5

Example E2.6

Draw a tree diagram for 3 tosses of a balanced (fair) coin. List all outcomes for this experiment in a sample space S.

SOLUTION:

Let H = a toss results in head

T = a toss results in tail.

Hence, the sample space is written as $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$.

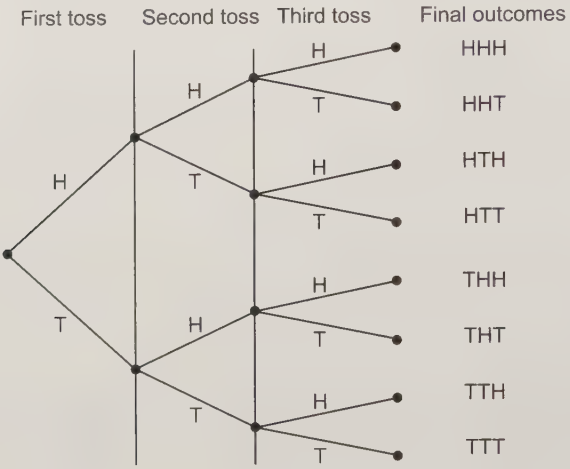


Fig. E2.6: Tree diagram

Example E2.7

Refer to Example E2.5. List all the outcomes included in each of the following events. Indicate which are simple and which are compound events.

- (a) both balls are different colours
- (b) at least one ball is red
- (c) not more than one ball is blue
- (d) the first ball is blue and the second is red.

SOLUTION:

- (a) $\{RB, BR\}$; a compound event
- (b) $\{RR, RB, BR\}$; a compound event
- (c) $\{RR, RB, BR\}$; a compound event
- (d) $\{BR\}$; a simple event.

2.9 APPROACHES TO PROBABILITY

There are three conceptual approaches to probability. They are

1. classical probability,
2. the relative frequency concept of probability and
3. the subjective probability concept.

2.9.1 Classical Probability

Outcomes that have the same probability of occurrence are called *equally likely outcomes*.

Two or more outcomes or events that have the same probability of occurrence are said to be *equally likely outcomes* or *events*. If E is an event, then $P(E)$ stands for the probability that event E occurs.

$$P(E_i) = \frac{1}{\text{total number of outcomes for the experiment}} \tag{2.5}$$

$$P(A) = \frac{\text{number of outcomes favourable to } A}{\text{total number of outcomes for the experiment}} \tag{2.6}$$

2.9.2 Relative Frequency Concept of Probability

If an experiment is repeated n times and an event A is observed f times, then, according to the relative frequency concept of probability

$$P(A) = \frac{f}{n}$$

The relative frequencies are not probabilities but approximate probabilities.

The *law of large numbers* states that if an experiment is repeated again and again, the probability of an event obtained from the relative frequency approaches the actual or theoretical probability.

2.9.3 Subjective Probability

Subjective probability is assigned arbitrarily. It is the probability assigned to an event based on subjective judgement, experience, information, and belief.

2.9.4 Marginal Probability

Marginal probability is the probability of a single event without consideration of any other event. Marginal probability is also known as *simple probability*.

2.9.5 Conditional Probability

It is the probability that an event will occur given that another event has already occurred.

If A and B are two events, then the conditional probability of A and B is written as $P(A|B)$ and read as "the probability of A given that B has already occurred".

The conditional probability of an event is the probability that the event occurs under the assumption that another event has occurred. The probability that event B occurs given that event A has occur is called a *conditional probability*. It is denoted by the symbol $P(B/A)$, which is read as "the probability of B given A". A is called the given event.

Joint and Marginal Probabilities

Data obtained by observing values of one variable of a population are known as *univariate data*. Data obtained by observing two variables of a population are called *bivariate data*, and a frequency distribution for bivariate data is called a *contingency table* or *two-way table*.

Example E2.8

A small bin contains 50 rubber balls. Of them, 26 are red and 24 are green. If one ball is randomly selected out of this bin, what is the probability that this ball is

(*a*) red

(*b*) green

SOLUTION:

(*a*) $P(\text{ball selected is red}) = \dfrac{26}{50} = 0.52$

(*b*) $P(\text{ball selected is green}) = \dfrac{24}{50} = 0.48$

Example E2.9

A multiple-choice question in a test contains five answers. If a student chooses one answer based on "pure guess", what is the probability that this student answer is

(*a*) correct

(*b*) wrong

Does these probabilities add up to 1? If yes, why?

SOLUTION:

(*a*) $P(\text{student's answer is correct}) = \dfrac{1}{5} = 0.2$

(*b*) $P(\text{student's answer is wrong}) = \dfrac{4}{5} = 0.8$

Yes, these probabilities add up to 1 because this experiment has two and only two outcomes, and according to the second property of probability, the sum of their probabilities must be equal to 1.

Example E2.10

A sample of 1000 families showed that 35 of them own no automobiles, 207 own one automobile each, 377 own two automobiles each, 264 own 3 automobiles each and 117 own 4 or more automobiles each. Write a frequency distribution for this problem. If one family is chosen randomly from these 1000 families, find the probability that this family owns

(*a*) two automobiles

(*b*) four or more automobiles.

SOLUTION:

Refer to Table E2.10.

(*a*) $P(\text{family selected owns 2 automobiles}) = 0.377$

(*b*) $P(\text{family selected owns 4 or more automobiles}) = 0.117$

Table E2.10

Automobiles owned	Frequency	Relative frequency
0	35	0.035
1	207	0.207
2	377	0.377
3	264	0.264
4 or more	117	0.117
Total	1000	1.000

2.10 MUTUALLY EXCLUSIVE EVENTS

Events that cannot occur together are known as *mutually exclusive events*. Such events do not have any common outcomes. If two or more events are mutually exclusive, then at most one of them will occur every time we repeat the experiment. Hence, the occurrence of one event excludes the occurrence of the other event or events. For example, consider tossing a balanced coin twice. This experiment has 4 outcomes: *HH*, *HT*, *TH* and *TT*. These outcomes are mutually exclusive since one and only one of them will occur when we toss this coin twice.

Two or more events are *mutually exclusive events* if no two of them have outcomes in common. The Venn diagram shown in Fig. 2.4 show the difference between two events that are mutually exclusive and two events that are not mutually exclusive. Similarly Fig. 2.5 shows three mutually exclusive events and two cases of three events that are not mutually exclusive. Two events are said to be *independent*, if the occurrence of one does not affect the probability of the occurrence of the other. Thus, if A and B are *independent events*, then either $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$. If the occurrence of one event affects the probability of the occurrence of the other event, then the two events are said to be dependent events.

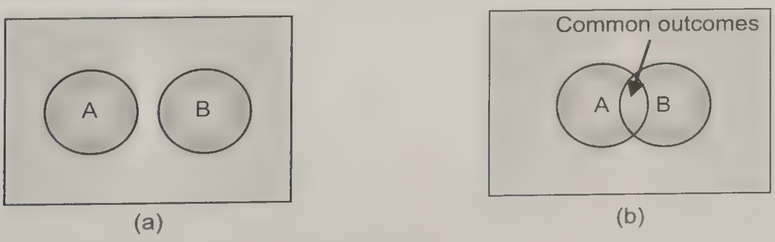
Hence, two events will be dependent if either $P(A \mid B) \neq P(A)$ or $P(B \mid A) \neq P(B)$. Two events are either *mutually exclusive* or *independent*. That is, mutually exclusive, events are always dependent, and dependent and independent events are never exclusive. Similarly, dependent events may or may not be mutually exclusive.

Two mutually exclusive events that are taken together include all the outcomes for an experiment are called *complementary event*. Thus, the complement of event A, denoted by $\bar{A}$ is the event that includes all the outcomes for an experiment that are not in A. It is clear that

$$P(A) + P(\bar{A}) = 1 \tag{2.7}$$

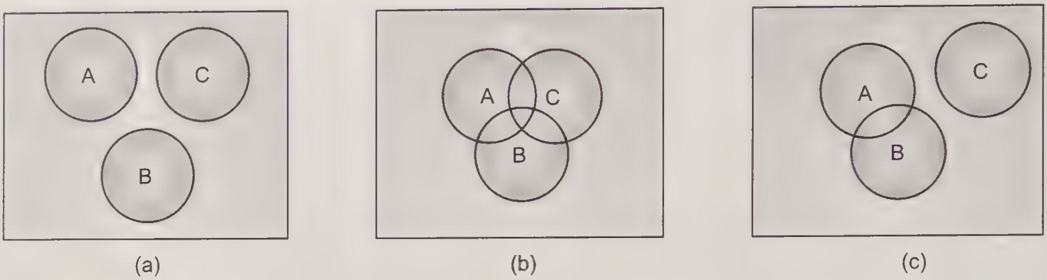
Also $$P(A) = 1 - P(\bar{A}) \text{ and } P(\bar{A}) = 1 - P(A) \tag{2.8}$$

The *intersection* of two events is given by the outcomes that are common to both events. If A and B are two events defined in a sample space, then, the intersection of A and B represent the collection of all outcomes that are common to both A and B and is denoted by A and B or AB or BA.



(a) (b)

Fig. 2.4



(a) (b) (c)

Fig. 2.5

2.11 INDEPENDENT AND DEPENDENT EVENTS

Two events are said to be *independent* if the occurrence of one does not affect the probability of the occurrence of the other. In other words, A and B are *independent events* if either

$$P(A \mid B) = P(A) \text{ or } P(B \mid A) = P(B) \tag{2.9}$$

It can be shown that if one of these two conditions is true, then the second will also be true, and if one is not true then the second will also not be true.

If the occurrence of one event affects the probability of the occurrence of the other event, then the two events are said to be *dependent events*.

Using the probability notation, the two events will be *dependent* if either

$$P(A \mid B) \neq P(A) \tag{2.10}$$

or

$$P(B \mid A) \neq P(B) \tag{2.11}$$

2.12 COMPLEMENTARY EVENTS

The complement event A denoted by $\bar{A}$ and read as "A bar" or "A complement" is the event that includes all the outcomes for an experiment that are not in A.

Events A and $\bar{A}$ are complements of each other. The Venn diagram in Fig. 2.6 shows the complementary events A and $\bar{A}$.

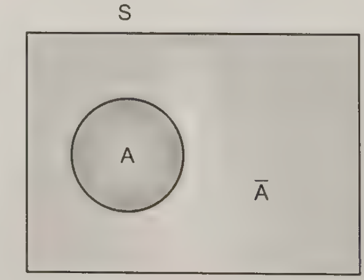


Fig. 2.6: Venn diagram of two complementary events

Since two complementary events are taken together, they include all the outcomes for an experiment and because the sum of the probabilities of all outcomes is 1, it is clear that

$$P(A) + P(\bar{A}) = 1 \tag{2.12}$$

Also $$P(A) = 1 - P(\bar{A}) \tag{2.13}$$

and $$P(\bar{A}) = 1 - P(A) \tag{2.14}$$

Example E2.11

A group of 2000 randomly selected adults were asked if they are in favour or against building a nuclear power plant in the city. The following table gives the results of this survey.

	In favour	Against	Total
Male	500	400	900
Female	650	550	1100
Total	1150	950	2000

(a) If one person is selected at random from these 2000 adults, find the probability that this person is

(i) in favour of the plant

(ii) against the plant

(iii) in favour of the plant given the person is a female

(iv) a male given the person is against the plant

(b) Are the events "male" and "in favour" mutually exclusive? What about the events "in favour" and "against"? Why or why not?

(c) Are the events "female" and "male" independent? Why or why not?

SOLUTION:

(a) *(i)* $P(\text{in favour}) = \dfrac{1150}{2000} = 0.575$

(ii) $P(\text{against}) = \dfrac{950}{2000} = 0.475$

(iii) $P(\text{in favour}|\text{female}) = \dfrac{650}{1100} = 0.59091$

(iv) $P(\text{male}|\text{against}) = \dfrac{400}{950} = 0.42105$

(b) The events "male" and "in favour" are not mutually exclusive because they can occur together. The events "in favour" and "against" are mutually exclusive because they cannot occur together.

(c) $P(\text{female}) = \dfrac{1100}{2000} = 0.55$

$P(\text{female}|\text{in favour}) = \dfrac{650}{1150} = 0.56522$

Since these two probabilities are not equal, the events "female" and "in favour" are not independent.

Example E2.12

There are a total of 170 practicing physicians in a big city. Of them, 60 are female and 30 are pediatricians. Of the 60 females, 10 are pediatricians. Are the events "female" and "pediatrician" independent? Are they mutually exclusive? Explain why or why not.

SOLUTION:

$$P(\text{pediatrician}) = \frac{30}{170} = 0.17647$$

$$P(\text{pediatrician}|\text{female}) = \frac{10}{60} = 0.16667$$

Since these two probabilities are not equal, the events "female" and "pediatrician" are not independent. The events are not mutually exclusive because they can occur together.

Example E2.13

A company hired 50 new graduates last month. Of these 25 are male and 15 are business majors. Of the 25 males, 15 are business majors. Are the events "male" and "business major" independent? Are they mutually exclusive? Explain why or why not.

SOLUTION:

$$P(\text{business major}) = \frac{15}{50} = 0.30$$

$$P(\text{business major}|\text{male}) = \frac{15}{25} = 0.60$$

Since these two probabilities are not equal, the events "male" and "business major" are not independent. The events are not mutually exclusive because they can occur together.

Example E2.14

Let A be the event that a number less than 3 is obtained if a die is rolled once. What is the probability of A? What is the complementary event of A, and what is its probability?

SOLUTION:

Event A will occur if either 1-spot or a 2-spot is obtained on the die. Hence,

$$P(A) = \frac{2}{6} = 0.3334$$

The complementary event of A is that either a 3-spot or 4-spot or a 5-spot, or a 6-spot is obtained on the die. Hence,

$$P(\bar{A}) = 1 - 0.3334 = 0.6666$$

2.13 INTERSECTION OF EVENTS AND MULTIPLICATION RULE

In this section, we introduce the intersection of two events and the application of multiplication rule to find the probability of the intersection of events.

2.13.1 Intersection of Events

Let A and B be two events defined in a sample space. The *intersection* of A and B represents the collection of all outcomes that are common to both A and B and is denoted by A and B. The intersection of events A and B is also denoted by either $A \cap B$ or AB. Figure 2.7 shows the intersection of events A and B. The shaded area in Fig. 2.7 gives the intersection of events A and B.

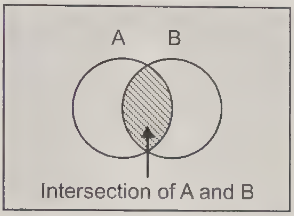


Fig. 2.7: Intersection of events A and B

2.13.2 Multiplication Rule

The probability that events A and B happen together is called the *joint probability* of A and B and is written as $P(A \text{ and } B)$. The probability of the intersection of two events is obtained by multiplying the marginal probability of one event by the conditional probability of second event. The rule is called the *multiplication rule*. Hence,

$$P(A \text{ and } B) = P(A)\, P(B|A) \tag{2.15}$$

The joint probability of events A and B can also be denoted by $P(A \cap B)$ or $P(AB)$.

Hence, for any two events, their joint probability is equal to the probability that one of the events occurs times the conditional probability of the other given event.

Conditional Probability

If A and B are two events, then,

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)} \tag{2.16}$$

and
$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \tag{2.17}$$

given that $P(A) \neq 0$ and $P(B) \neq 0$.

Hence, for any two events A and B, the conditional probability that one event occurs given that the other event has occurred equals the joint probability of the two events divided by the probability of the given event.

Multiplication Rule for Independent Events

The probability of the intersection of two independent events A and B is

$$P(A \text{ and } B) = P(A)\, P(B) \tag{2.18}$$

The Special Multiplication Rule

The general multiplication rule for any two events A and B is given by Eq. (2.16)

$$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$$

If A and B are independent events, then

$$P(B \mid A) = P(B) \tag{2.19}$$

Hence, for the special case of independent events, we replace the term $P(B \mid A)$ in the general multiplication rule by the term $P(B)$.

If A and B are independent events, then

$$P(A \text{ and } B) = P(A) \cdot P(B) \tag{2.20}$$

and conversely, if

$$P(A \text{ and } B) = P(A)\, P(B)$$

then A and B are independent events. Hence, two events are independent if and only if their joint probability equals the product of their marginal probabilities.

Similarly, if events A, B, C,... are independent, then

$$P(A \text{ and } B \text{ and } C \text{ and } \ldots) = P(A) \cdot P(B) \cdot P(C) \ldots \tag{2.21}$$

Mutually exclusive events are those that cannot occur simultaneously. *Independent events* are those for which the occurrence of some does not affect the probabilities of the others occurring. Two or more (non-impossible) events cannot be both mutually exclusive and independent.

Total Probability Rule

Events $A_1, A_2, A_3, \ldots, A_k$ are said to be *exhaustive*, if one or more of them must occur. Suppose events $A_1, A_2, A_3, \ldots, A_k$ are mutually exclusive and exhaustive; that is exactly one of the events must occur. Then for any event B

$$P(B) = \sum_{j=1}^{k} P(A_j) \cdot P(B \mid A_j) \tag{2.22}$$

Joint Probability of Mutually Exclusive Events

The joint probability of two mutually exclusive events is always 0. If A and B are two mutually exclusive events, then

$$P(A \text{ and } B) = 0 \tag{2.23}$$

Example E2.15

In a group of 10 adults, 4 have type A personality and 6 have a type B personality. If two persons are selected at random from this group of 10, what is the probability that the first of them has a type A personality and the second has a type B personality? Draw a tree diagram for this experiment.

SOLUTION:

Let C = first person selected has a type A personality

D = first person selected has a type B personality

E = second person selected has a type A personality

F = second person selected has a type B personality

Figure E2.15 shows the tree diagram for the experiment.

The probability that the first person has a type A personality and the second has a type B personality is given by

$$P(C \text{ and } F) = P(C)\, P(F|C) = \left(\frac{4}{10}\right)\left(\frac{6}{9}\right) = 0.2667.$$

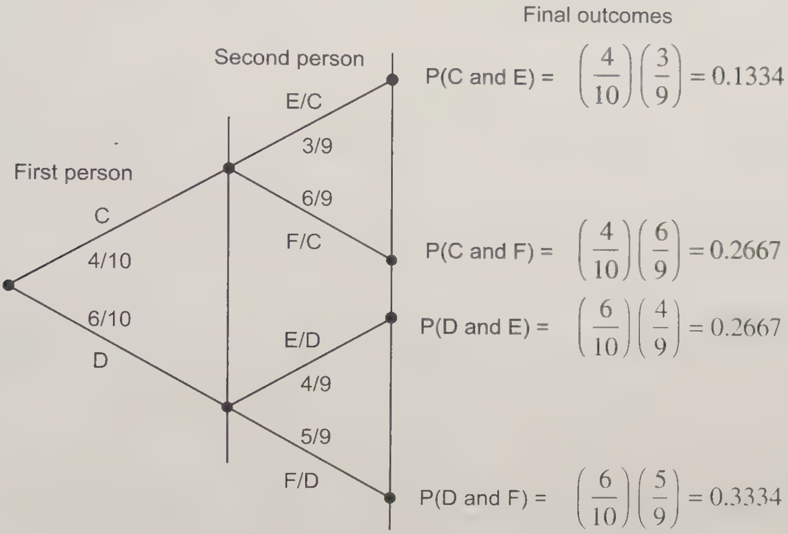

Final outcomes
Second person
First person
C
4/10
6/10
D
E/C
3/9
6/9
F/C
E/D
4/9
5/9
F/D
P(C and E) = $\left(\frac{4}{10}\right)\left(\frac{3}{9}\right) = 0.1334$
P(C and F) = $\left(\frac{4}{10}\right)\left(\frac{6}{9}\right) = 0.2667$
P(D and E) = $\left(\frac{6}{10}\right)\left(\frac{4}{9}\right) = 0.2667$
P(D and F) = $\left(\frac{6}{10}\right)\left(\frac{5}{9}\right) = 0.3334$

Fig. E2.15: Tree diagram

Example E2.16

Ten per cent of all items sold by a mail order company are returned by customers for a refund. Find the probability that of two items sold during a given hour by this company.

(a) both will be returned for a refund

(b) neither will be returned for a refund

Draw a tree diagram for this experiment.

SOLUTION:

Let A = first item is returned

B = first item is not returned

C = second item is returned

D = second item is not returned

(a) $P(A \text{ and } C) = P(A)\, P(C) = (0.10)(0.10) = 0.01$

(b) $P(B \text{ and } D) = P(B)\, P(D) = (0.90)(0.90) = 0.81$

The tree diagram for this equipment is shown in Fig. E2.16.

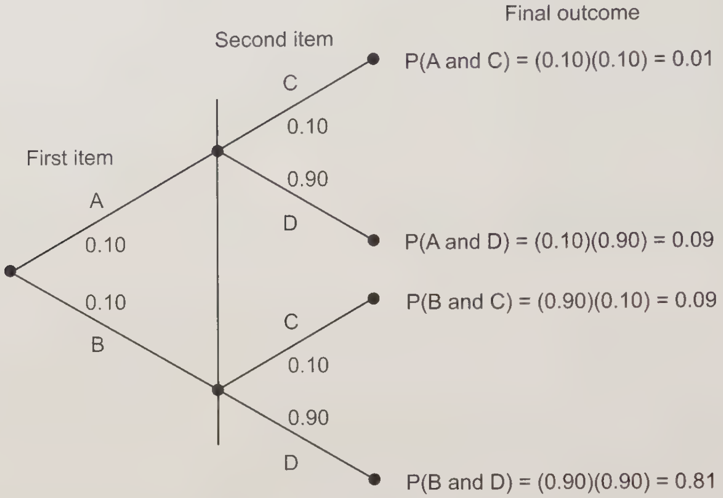

Final outcome
Second item
C
0.10
0.90
D
First item
A
0.10
0.10
B
C
0.10
0.90
D
P(A and C) = (0.10)(0.10) = 0.01
P(A and D) = (0.10)(0.90) = 0.09
P(B and C) = (0.90)(0.10) = 0.09
P(B and D) = (0.90)(0.90) = 0.81

Fig. E2.16: Tree diagram

Example E2.17

The probability that a farmer is in debt is 0.80. What is the probability that three randomly selected farmers are all in debt? Assume independence of events.

SOLUTION:

Let D_1 = the first farmer is in debt

D_2 = the second farmer is in debt

D_3 = the third farmer is in debt

Hence, $P(D_1 \text{ and } D_2 \text{ and } D_3) = P(D_1)P(D_2)P(D_3) = (0.80)(0.80)(0.80) = 0.512$

Example E2.18

According to some private estimates, the probability that a randomly selected student from the population of all students enrolled in all institutions of higher education in a particular state is a female is 0.60, and the probability that this student is a female and a part-time student is 0.25. What is the probability that this student is part-time given that the student is a female?

SOLUTION:

Let F = student selected is a female

PT = student selected is a part-time

It is given that

$P(F) = 0.60$ and $P(PT \text{ and } F) = 0.25$

Hence, $P(PT \mid F) = \dfrac{P(PT \text{ and } F)}{P(F)} = \dfrac{0.25}{0.60} = 0.41667$

2.14 UNION OF EVENTS AND THE ADDITION RULE

In this section, we will discuss the union of events and the addition rule to compute the probability of the union of events.

2.14.1 Union of Events

The *union of two events* A and B includes all outcomes that are either in A or in B or in both A and B. Let A and B be two events defined in a sample space. The *union of events* A and B is the collection of all outcomes that belong either to A or to B or to both A and B and is denoted by A or B.

The union of events A and B is also denoted by "$A \cup B$".

2.14.2 Addition Rule

The method used to calculate the probability of the union of events is called the addition rule.

The probability of the union of two events A and B is

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

where $P(A)$ and $P(B)$ are the marginal probabilities of A and B respectively and $P(A$ and $B)$ is the joint probability of A and B.

Addition Rule for Mutually Exclusive Events

The probability of the union of two mutually exclusive events A and B is

$$P(A \text{ or } B) = P(A) + P(B) \tag{2.24}$$

Special Addition Rule

If event A and event B are mutually exclusive, then

$$P(A \text{ or } B\,) = P(A) + P(B) \tag{2.25}$$

Generalizing, if events $A, B, C, \ldots$ are mutually exclusive, then

$$P(A \text{ or } B \text{ or } C \text{ or } \ldots) = P\,(A) + P\,(B) + P(C) + \cdots \tag{2.26}$$

Therefore, for mutually exclusive events, the probability that one or another of the events occurs equals to the sum of the individual probabilities.

The General Addition Rule

For events that are not mutually exclusive, the general addition rule is applied. Referring to Fig. 2.8, we see that

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \tag{2.27}$$

For any two events, the probability that one or the other occurs equals the sum of the individual probabilities less the probability that both occur.

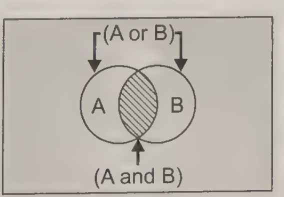


Fig. 2.8: Non-mutually exclusive events

Example E2.19

All 500 employees of a company were asked whether they are smokers or non-smokers of cigarettes and whether or not they are college graduates. Based on this information, the following two-way classification table was prepared:

	College graduate	Not a college graduate	Total
Smoker	55	100	155
Non-smoker	150	195	345
Total	205	295	500

Suppose one employee is selected at random from this company. Find the following probabilities:

(a) P(college graduate or smoker)

(b) P(smoker or not a college graduate)

(c) P(smoker or non-smoker)

SOLUTION:

Let S = smoker

N = non-smoker

C = college graduate

D = not a college graduate

(a) $P(C \text{ or } S) = P(C) + P(S) - P(C \text{ and } S) = \dfrac{205}{500} + \dfrac{155}{500} - \dfrac{55}{500} = 0.60$

(b) $P(S \text{ or } D) = P(S) + P(D) - P(S \text{ and } D) = \dfrac{155}{500} + \dfrac{295}{500} - \dfrac{100}{500} = \dfrac{350}{500} = 0.7$

(c) Since S and N are mutually exclusive events

$$P(S \text{ or } N) = P(S) + P(N) = \frac{155}{500} + \frac{345}{500} = 1.0$$

Example E2.20

The probability that a family owns a desktop computer is 0.78 that it owns a DVD player is 0.70, and that it owns both the computer and a DVD is 0.60. Find the probability that a randomly selected family owns a computer or a DVD player.

SOLUTION:

Let C = family selected owns a computer

D = family selected owns a DVD player

Then, $P(C \text{ or } D) = P(C) + P(D) - P(C \text{ and } D) = 0.78 + 0.70 - 0.60 = 0.88$

Example E2.21

The probability that an open-heart surgery is successful is 0.85. What is the probability that in two randomly selected open-heart surgeries at least one will be successful? Draw a tree diagram for this experiment.

SOLUTION:

Let A = first open-heart surgery is successful

B = first open-heart surgery is not successful

C = second open-heart surgery is successful

D = second open-heart surgery is not successful

P(at least one open-heart surgery is successful) = $P(A \text{ and } C) + P(A \text{ and } D) + P(B \text{ and } C)$

$= (0.85)(0.85) + (0.85)(0.15) + (0.15)(0.85) = 0.7225 + 0.1275 + 0.1275 = 0.9775$

The tree diagram is shown in Fig. E2.21.

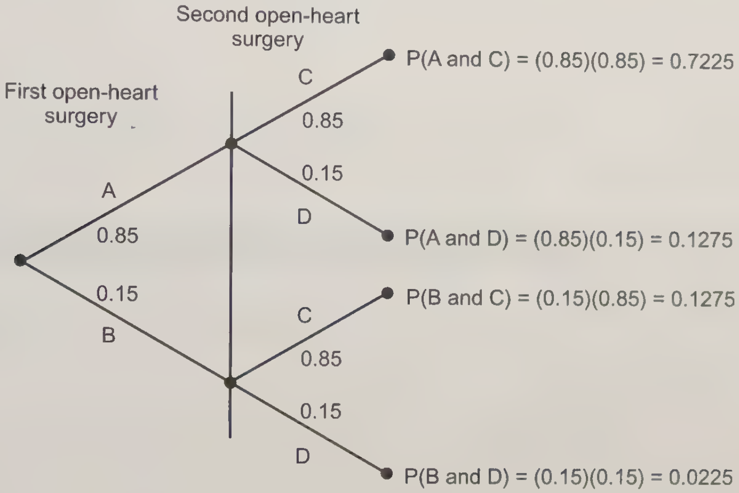

Second open-heart surgery
First open-heart surgery
A
0.85
0.15
B
C
0.85
0.15
D
C
0.85
0.15
D
P(A and C) = (0.85)(0.85) = 0.7225
P(A and D) = (0.85)(0.15) = 0.1275
P(B and C) = (0.15)(0.85) = 0.1275
P(B and D) = (0.15)(0.15) = 0.0225

Fig. E2.21

Example E2.22

(a) In a class of 35 students, 11 are seniors, 8 are juniors, 9 are sophomores, and 7 are freshmen. If one student is selected at random from this class, what is the probability that this student is

(i) a junior

(ii) a freshman

(b) If two students are selected at random from this class of 35 students, find the probability that the first student selected is a junior and the second is a sophomore.

SOLUTION:

(a) *(i)* $P(\text{student selected is a junior}) = \dfrac{8}{35} = 0.22857$

(ii) $P(\text{student selected is a freshman}) = \dfrac{7}{35} = 0.2$

(b) Let J_1 = first student selected is a junior

S_2 = second student selected is a sophomore

$$P(J_1 \text{ and } S_2) = P(J_1)\,P(S_2|J_1) = \left(\frac{8}{35}\right)\left(\frac{9}{35}\right) = 0.05878$$

Example E2.23

A manufacturing company makes automobile steering wheels. The manufacturing system involves two independent processing machines so that each steering wheel passes through these two processes. The probability that the first processing machine is not working properly at any time is 0.09, and the probability that the second machine is not working properly at any time is 0.065. Find the probability that both machines will not be working properly at any given time.

SOLUTION:

$$\begin{aligned} &P(\text{both machines not working properly}) \\ &= P(\text{first machine is not working properly})\,P(\text{second machine not working properly}) \\ &= (0.09)(0.065) = 0.00585 \end{aligned}$$

2.15 BAYE'S FORMULA

Baye's formula is quite useful in modifying a probability estimate as additional information becomes available. We have seen that a conditional probability is one in which the probability of the event depends upon whether the other event has occurred.

$$P(A/B) = \frac{P(AB)}{P(B)} \tag{2.28}$$

or

$$P(B/A) = \frac{P(AB)}{P(A)} \tag{2.29}$$

From Eqs. (2.28) and (2.29), we have

$$P(AB) = P(A/B)\,P(B) = P(B/A)\,P(A) \tag{2.30}$$

since

$$AB = BA$$

and

$$P(A/B) = \frac{P(A/B)\,P(A)}{P(B)} \tag{2.31}$$

Since $P(A) + P(\bar{A}) = 1$, it follows that event B occurs jointly with either A or $\bar{A}$.

$$P(B) = P(AB) + P(\bar{A}B)$$

and from Eq. (2.30)

$$P(B) = P(A)\,P(B/A) + P(\bar{A})\,P(B/\bar{A}) \tag{2.32}$$

Substituting Eq. (2.32) into Eq. (2.31), we get

$$P(A/B) = \frac{P(A)\,P(B/A)}{P(A)\,P(B/A) + P(\bar{A})\,P(B/\bar{A})} \tag{2.33}$$

If event A has more than two available alternatives, we can write Eq. (2.33) as

$$P(A/B) = \frac{P(A_i)\,P(B/A_i)}{\sum_i P(A_i)\,P(B/A_i)} \tag{2.34}$$

Example E2.24

Find the joint probability of A and B for the following:

(a) $P(B) = 0.60$ and $P(A/B) = 0.70$

(b) $P(A) = 0.2$ and $P(B/A) = 0.15$

SOLUTION:

(a) $P(A \text{ and } B) = P(B \text{ and } A) = P(B)\,P(A/B) = (0.60)(0.70) = 0.42$

(b) $P(A \text{ and } B) = P(A)\,P(B/A) = (0.2)(0.15) = 0.03$

Example E2.25

A certain manufacturing company has 456 employees. Based on the information obtained by asking all the employees whether they are smokers or non-smokers and whether or not they are married, the following two-way classification table of data was obtained.

Smoker/Non-smoker	Married	Not married	Total
Smoker	37	72	109
Non-smoker	121	226	347
Total	158	298	456

(a) If one employee is selected at random from this company, find the probabilities of the following:

(i) P(Married and non-smoker)

(ii) P(Smoker and not married)

(b) Draw a tree diagram and compute all the joint probabilities for the given data.

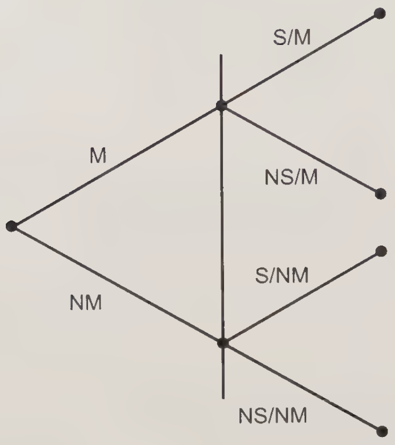
SOLUTION:

(a) Let
M = Married
NM = Not married
S = Smoker
NS = Non-smoker

(i) $P(M \text{ and } NS) = P(M)\,P(NS/M) = \left(\frac{158}{456}\right)\left(\frac{121}{158}\right) = 0.2653$

(ii) $P(S \text{ and } NM) = P(S)\,P(NM/S) = \left(\frac{109}{456}\right)\left(\frac{72}{109}\right) = 0.1579$

(b) The tree diagram is shown in Fig. E2.25.



$$P(\text{M and S}) = \left(\frac{158}{456}\right)\left(\frac{37}{158}\right) = 0.0811$$

$$P(\text{M and NS}) = \left(\frac{158}{456}\right)\left(\frac{121}{158}\right) = 0.2653$$

$$P(\text{NM and S}) = \left(\frac{298}{456}\right)\left(\frac{72}{298}\right) = 0.1579$$

$$P(\text{NM and NS}) = \left(\frac{298}{456}\right)\left(\frac{226}{298}\right) = 0.4956$$

Fig. E2.25: Tree diagram

Example E2.26

Refer to the data of Example E2.25, if one employee is selected at random from the company, calculate the following probabilities.

(*a*) *P*(married or smoker)

(*b*) *P*(smoker or not married)

(*c*) *P*(smoker or non-smoker)

SOLUTION:

Referring to the solution of Example E2.25, we have

(*a*) $P(M \text{ or } S) = P(M) + P(S) - P(M \text{ and } S) = \left(\frac{158}{456}\right) + \left(\frac{109}{456}\right) - \left(\frac{37}{456}\right) = 0.5044$

(*b*) $P(S \text{ or } NM) = P(S) + P(NM) - P(S \text{ and } NM) = \left(\frac{109}{456}\right) + \left(\frac{298}{456}\right) - \left(\frac{72}{456}\right) = 0.7346$

(*c*) Since S and NS are mutually exclusive events,

$$P(S \text{ or } NS) = P(S) + P(NS) = \left(\frac{109}{456}\right) + \left(\frac{347}{456}\right) = 1.0$$

Example E2.27

The probability that a manufacturer is in a debt is 0.7. What is the probability that three randomly selected manufacturers are all in debt? Assume independence of events.

SOLUTION:

Let M_1 = first manufacturer in debt

M_2 = second manufacturer in debt

M_3 = third manufacturer in debt

Then, $P(M_1 \text{ and } M_2 \text{ and } M_3) = P(M_1)\,P(M_2)\,P(M_3) = (0.7)(0.7)(0.7) = 0.343$

Example E2.28

Find $P(A \text{ or } B)$ for the following:

(a) $P(A) = 0.17$, $P(B) = 0.5$ and $P(A \text{ and } B) = 0.12$

(b) $P(A) = 0.8$, $P(B) = 0.7$ and $P(A \text{ and } B) = 0.6$.

SOLUTION:

(a) $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.17 + 0.5 - 0.12 = 0.55$

(b) $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) = 0.8 + 0.7 - 0.6 = 0.9$

Example E2.29

The average acceptability of three automobile transmissions from three suppliers is given below:

Supplier	Relative probability of supply	Average acceptability
1	0.56	0.97
2	0.32	0.89
3	0.12	0.77

(a) Find the probability that any transmission received will perform acceptably

(b) Find the probability that a particular transmission delivered by supplier 2 when it is known to have performed acceptably.

SOLUTION:

(a) $P(A) = (0.56)(0.97) + (0.32)(0.89) + (0.12)(0.77) = 0.5432 + 0.2848 + 0.0924 = 0.9204$

(b)
$$P(T_2 / A) = \frac{P(T_2)P(A/T_2)}{\Sigma P(T_1)P(A/T_1)} = \frac{(0.32)(0.89)}{(0.56)(0.97)+(0.32)(0.89)+(0.12)(0.77)} = \frac{0.2848}{0.9204} = 0.3094$$

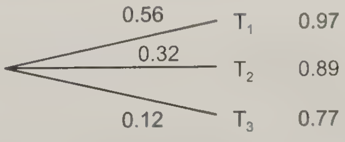


Fig. E2.29

2.16 ADDITIONAL EXAMPLES AND SOLUTIONS

Example E2.30

Following is a percentage distribution for the number of years of school completed by adult employees, 25 years old or over in a manufacturing company.

Years completed	Percentage	Event
0–4	2	A
5–7	4	B
8	5	C
9–11	11	D
12	40	E
13–15	20	F
16 or more	18	G

Determine the probability that an employee randomly selected

(a) has completed 8 years of school or less

(b) has completed 12 years of school or less

(c) has completed 13 years of school or more

(d) interpret each of your answers in *(a)–(c)* in terms of percentages.

SOLUTION:

(a) $P(\text{the employee has completed 8 years of school or less}) = P(A) + P(B) + P(C)$
$= 0.02 + 0.04 + 0.05 = 0.11$

(b) $P(\text{the employee has completed 12 years of school or less}) = P(A) + P(B) + P(C) + P(D) + P(E)$
$= 0.02 + 0.04 + 0.05 + 0.11 + 0.40 = 0.62$

(c) $P(\text{the employee has completed 13 years of school or more}) = P(F) + P(G)$
$= 0.20 + 0.18 = 0.38$

(d) The interpretation of each of the results above in terms of percentage is as follows:

(i) 11% of the adults aged 25 years old and over have completed 8 years of school or less

(ii) 62% of the adults aged 25 years old and over have completed 12 years of school or less

(iii) 38% of the adults aged 25 years old and over have completed 13 years of school or more.

Example E2.31

In a particular corporation, 56% of the employees are white, 95% are male, and 53% are white males. For a randomly selected employee, let

W = Event the employee selected is white
M = Event the employee selected male

(a) Find $P(W)$, $P(M)$ and $P(W \text{ and } M)$

(b) Determine $P(W \text{ or } M)$ and express the answer in percentages

(c) Find the probability that a randomly selected employee is female.

SOLUTION:

(a) $P(W) = 0.56$
$P(M) = 0.95$
$P(W \text{ or } M) = 0.53$

(b) $P(W \text{ or } M) = 0.56 + 0.95 - 0.53 = 0.98$
98% of the employees are either white or male.

(c) $P(F) = 1 - 0.95 = 0.05$

Example E2.32

The number of two door cars and four door vehicles in use by the employees of a company by age are as shown in the following Table E2.32.

Table E2.32

Age (years)		Type		Total
		2 Door T	4 Door F	
Under 6	A_1	25	45	70
6–8	A_2	15	5	20
9–11	A_3	20	10	30
12 and over	A_4	13	7	20
Total		73	67	140

For a randomly selected vehicle (2 Door or 4 Door), determine the probability that the vehicle selected

(a) is under 6 years old
(b) is under 6 years old, given that it is a 2 Door
(c) is a 2 Door
(d) is a 2 Door; given that it is under 6 years old
(e) interpret your answers in *(a)* to *(c)* in terms of percentages.

SOLUTION:

The probability that the vehicle selected is

(a) Under 6 years old is 70/140 = 0.5
(b) Under 6 years old, given that it is a 2 Door car is 25/73 = 0.343
(c) A 2 door car is 73/140 = 0.521
(d) A 2 door car; given that it is under 6 years old is 25/70 = 0.357
(e) *(i)* 50% of all vehicles are under 6 years old
(ii) 34.3% of all 2 door cars are under 6 years old
(iii) 52.1% of all vehicles are 2 door
(iv) 35.7% of all vehicles under 6 years old are 2 door cars.

Example E2.33

Table E2.33 presents a joint probability distribution for engineers and scientists by highest degree obtained in a particular *R&D* corporation.

Table E2.33

Highest degree	Type		
	Engineer, T_1	Scientist, T_2	$P(D_1)$
Bachelors, D_1	0.35	0.29	0.64
Masters, D_2	0.10	0.2	0.30
Doctorate, D_3	0.02	0.025	0.045
Other, D_4	0.01	0.005	0.015
$P(T_1)$	0.48	0.52	1.0

Determine the probability that the person selected

- *(a)* is an engineer
- *(b)* has a doctorate
- *(c)* is an engineer with a doctorate
- *(d)* is an engineer given the person has a doctorate
- *(e)* has a doctorate, given the person is an engineer
- *(f)* interpret your answers in *(a)–(e)* in terms of percentages.

SOLUTION:

The probability that the person selected

- (a) is an engineer is 0.48.
- *(b)* has a doctorate is 0.045.
- *(c)* is an engineer with a doctorate is 0.02.
- *(d)* is an engineer given the person has a doctorate is 0.02/0.045 = 0.444.
- *(e)* has a doctorate, given that the person is an engineer is 0.02/0.48 = 0.0417.
- *(f)* *(i)* 48% of all engineers and scientists are engineers.
  - *(ii)* 4.5% of all engineers and scientists have doctorates.
  - *(iii)* 2% of all engineers and scientists are engineers with a doctorate.
  - *(iv)* 44.4% of all engineers and scientists with doctorates are engineers.
  - *(v)* 4.17% of all engineers have doctorates.

Example E2.34

A frequency distribution for the class level of students in "Introduction to JAVA programming" course is as given in Table E2.34.

Table E2.34

Class	Frequency
Freshman	5
Sophomore	14
Junior	11
Senior	6
Total	36

Two students are randomly selected without replacement. Find the probability that

- *(a)* the first student selected is a junior and the second a senior
- *(b)* both students selected are sophomores
- *(c)* draw a tree diagram for the solution of all the probabilities
- *(d)* what is the probability that one of the students selected is a freshman and the other a sophomore?

SOLUTION:

Let F_i = Freshman on selection i, where $i = 1, 2$

S_i = Sophomore on selection i, where $i = 1, 2$

J_i = Junior on selection i, where $i = 1, 2$

H_i = Senior on selection i, where $i = 1, 2$

(a) $P(J_1 \text{ and } H_2) = (11/36)/(6/35) = 0.0524$

(b) $P(S_1 \text{ and } S_2) = (14/36)/(13/35) = 0.1444$

(c) The tree diagram is shown in Fig. E2.34.

(d) The probability that one of the students selected is a freshman and the other student selected is a sophomore is:

$$P(F_1 \text{ and } S_2) + P(S_1 \text{ and } F_2) = 0.0555 + 0.0555 = 0.111$$

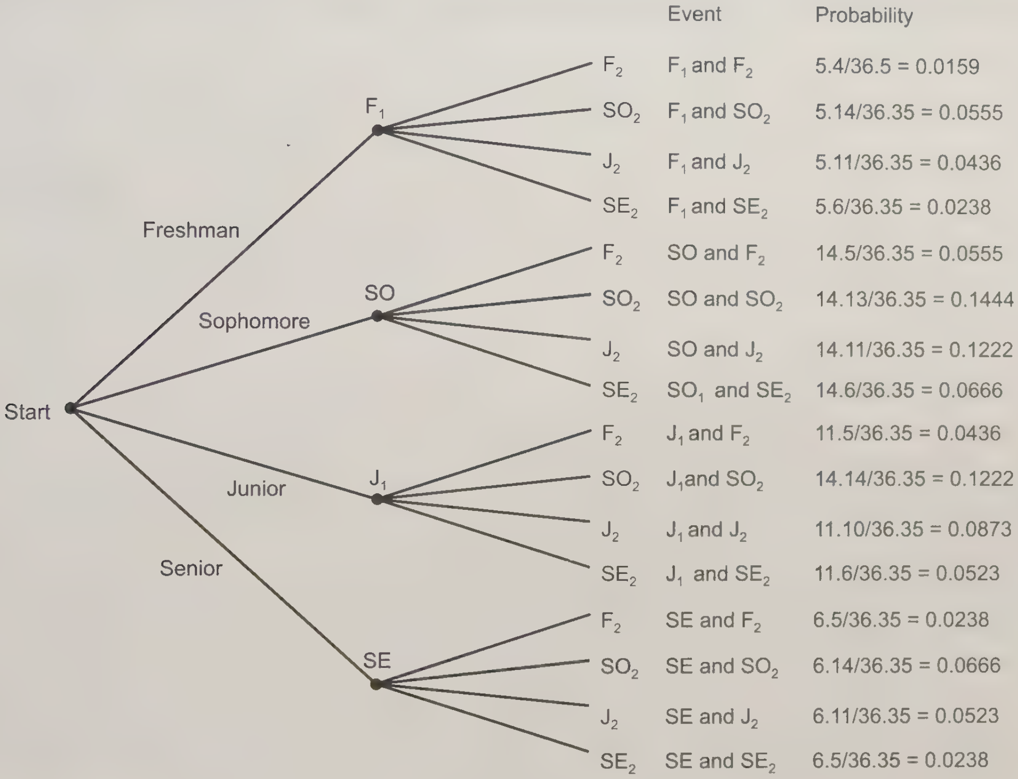

Event
Probability
Start
Freshman
Sophomore
Junior
Senior
F1
SO
J1
SE
F2 | F1 and F2 | 5.4/36.5 = 0.0159
SO2 | F1 and SO2 | 5.14/36.35 = 0.0555
J2 | F1 and J2 | 5.11/36.35 = 0.0436
SE2 | F1 and SE2 | 5.6/36.35 = 0.0238
F2 | SO and F2 | 14.5/36.35 = 0.0555
SO2 | SO and SO2 | 14.13/36.35 = 0.1444
J2 | SO and J2 | 14.11/36.35 = 0.1222
SE2 | SO1 and SE2 | 14.6/36.35 = 0.0666
F2 | J1 and F2 | 11.5/36.35 = 0.0436
SO2 | J1and SO2 | 14.14/36.35 = 0.1222
J2 | J1 and J2 | 11.10/36.35 = 0.0873
SE2 | J1 and SE2 | 11.6/36.35 = 0.0523
F2 | SE and F2 | 6.5/36.35 = 0.0238
SO2 | SE and SO2 | 6.14/36.35 = 0.0666
J2 | SE and J2 | 6.11/36.35 = 0.0523
SE2 | SE and SE2 | 6.5/36.35 = 0.0238

Fig. E2.34: Tree diagram

Example E2.35

Refer to Example E2.33

(a) determine $P(T_2)$, $P(D_3)$ and $P(T_2 \text{ and } D_3)$

(b) are T_2 and D_3 independent events? Why?

SOLUTION:

(a) $P(T_2) = 0.52$

$P(D_3) = 0.045$

$P(T_2 \text{ and } D_3) = 0.025$

(b) The special multiplication rule states that two events T_2 and D_3 are independent if,

$P(T_2 \text{ and } D_3) = P(T_2) \cdot P(D_3)$

Using the data in *(a)*

$P(T_2) \cdot P(D_3) = (0.52)(0.045) = 0.0234$

Since this product does not equal $P(T_2 \text{ and } D_3) = 0.025$, the events T_2 and D_3 are not independent.

Example E2.36

The two-way classification of data of 2000 randomly selected employees of a manufacturing company from a city based on gender and commuting time from residence to work place is given in Table E2.36.

Table E2.36

Gender	Commuting time from residence to work place			Total
	Less than 30 min	30 min to 1 hour	More than 1 hour	
Men	525	450	225	1200
Women	410	250	140	800
Total	935	700	365	2000

(a) If one employee is selected at random from these 2000 employees, find the probability that this employee
  - *(i)* commutes from more than 1 hour
  - *(ii)* commute for less than 30 minutes
  - *(iii)* is a man commuting for 30 min to 1 hour
  - *(iv)* is a woman commuting for more than 1 hour.

(b) Determine if the following events are mutually exclusive
  - *(i)* 'men' and 'commuting for more than 1 hour'
  - *(ii)* 'less than 30 min' 'more than 1 hour'
  - *(iii)* 'woman' and 'commutes for 30 minutes to 1 hour'. Are these events independent?

SOLUTION:

(a) *(i)* $P(\text{commutes for more than 1 hour}) = \dfrac{365}{2000} = 0.1825$

(ii) $P(\text{commutes for less than 30 min.}) = \dfrac{935}{2000} = 0.4675$

(iii) $P(\text{commutes for 30 min to 1 hour}) = \dfrac{450}{700} = 0.6428$

(iv) $P(\text{commutes for more than 1 hour/woman}) = \dfrac{140}{800} = 0.175$

(b) *(i)* The term 'man' and 'commutes for more than 1 hour' are not mutually exclusive. The two events can occur together.

(ii) The events 'less than 30 minutes' and 'more than 1 hour' are mutually exclusive because they cannot occur together.

(*iii*) $P(\text{woman}) = \dfrac{800}{2000} = 0.4$

$$P(\text{woman/commutes for 30 min to 1 hour}) = \frac{250}{700} = 0.3571$$

Since these two probabilities are not equal, the events ‘woman’ and ‘commute for 30 minutes to 1 hour’ are not independent.

2.17 SUMMARY

The term probability is used to describe the uncertainty *probability*, which measures the likelihood that an event will occur, is an important part of statistics. Probability theory is used extensively in analysing decision-making situations that involve risk or incomplete information.

PROBLEMS

P2.1 In Fig. P2.1, *E*, *T*, and *N* are the events that an automobile brought to a service garage needs an engine oil-change, tune-up, or new tires. Express in words the events represented by

(*a*) region 1
(*b*) region 3
(*c*) region 7
(*d*) regions 1 and 4 together
(*e*) regions 2 and 5 together
(*f*) regions 3, 5, 6, and 8 together.

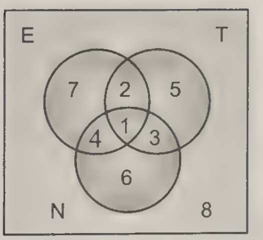


Fig. P2.1: Venn diagram

P2.2 In Problem 2.1, list the region or combinations of regions representing the events that an automobile brought to the garage needs

(*a*) tune-up, but neither engine oil-change nor new tires
(*b*) an engine oil-change and tune-up
(*c*) tune-up or new tires, but not an engine oil-change
(*d*) new tires.

P2.3 An electronic security system may be used in any one of the three modes, in any two of the three modes, or in all three modes: manual (*M*), semi-automatic (*S*) and automatic (*A*). In tests on 200 such systems, the following uses were found.

Number	Use
30	M∩S∩A
40	M∩S
20	A-only
40	S∩A
All remaining	M-only

(*a*) define *M*, *S*, and *A* as sets of systems used. Draw a Venn diagram, and show the number of systems in each use

(*b*) how many systems are in *M*? In *S*? In $(M \cup S)$?

P2.4 An age distribution of 100 employees at a manufacturing company are given as follows:

Age (years)	Number of employees
Under 40	1
40–49	14
50–59	41
60–69	27
70 and over	17
Total	100

For an employee selected at random, let

A : event the employee is under 40

B : event the employees is in his or here 40's

C : event the employee is in his or her 50's

S : event the employee is under 60

(*a*) use the tables and the f/n rule to find $P(S)$

(*b*) express event S in terms of events A, B and C

(*c*) determine $P(A)$, $P(B)$ and $P(C)$

(*d*) complete $P(S)$ using special addition rule and write the answers from parts (*b*) and (*c*). Also compare the answer in part (*a*).

P2.5 Refer to Problem P2.4. Find the probability that a randomly selected employee is

(*a*) 40 years old or under

(*b*) under 60 years old.

P2.6 In a particular company, 50% of the adult employees are female, 10% are divorced, and 5% are divorced females. For an employee selected at random, let

F = event the person is female

D = event the person is divorced.

(a) find $P(F)$, $P(D)$ and $P(F \text{ and } D)$

(b) determine $P(F \text{ or } D)$ and interpret your answer in terms of percentages

(c) find the probability that a randomly selected adult is male.

P2.7 A fair coin is tossed five times. Find the probability of getting at least one tail.

P2.8 Suppose we randomly select two students from a typical co-ed class and observe whether the student selected each time is a male or a female. Write all the outcomes for this experiment. Draw the Venn and tree diagrams for the experiment.

P2.9 In a group of freshman engineering class, some are in favour of mechanical engineering and others are against it. Two students are selected at random from this class and asked whether they are in favour of or against mechanical engineering. How many distinct outcomes are possible? Draw a Venn diagram and a tree diagram for this experiment. List all the outcomes included in each of the following events and state whether they are simple or compound events.

(a) both students are in favour of mechanical engineering

(b) at most one student is against mechanical engineering

(c) exactly one student in favour of mechanical engineering.

P2.10 Determine the conditional probability P(in favour/male) for the response data in Table P2.10 for the 100 employees. They were asked whether they are in favour or against paying higher salaries to CEOs of corporations.

Table P2.10

	In favour	Against	Total
Male	20	40	60
Female	4	36	40
Total	24	76	100

Draw the tree diagram to illustrate the computations.

P2.11 Refer to Problem P2.10. Find the conditional probability that a randomly selected employee is a female given that this employee is in favour of paying high salaries to the CEOs. Draw the tree diagram for this computation.

P2.12 *(a)* Determine the probability of getting 3 tails in tossing 3 coins.

(b) Two dice are thrown. Draw the sample space. Determine the probability of getting 6 in both the dice.

(c) Determine the probability of getting *(i)* exactly two tails (event A) and *(ii)* at least two tails (event B) in tossing 3 coins.

P2.13 *(a)* Determine the probability that the sum of the numbers shown on the two faces, when two dice are thrown

(i) is 7
(ii) is 10

(b) A die is thrown. Determine the probability of getting either an even number or a number greater than 4 or both.

P2.14 Table P2.14 gives the contingency table for age and rank of employees in a particular company. If an employee is selected at random, (*a*) find the unconditional probability that the employee selected is in his/her 50's

(*b*) find the unconditional probability that the employee selected is in his or her 50's given that Grade 1 employee is selected

(*c*) interpret the probabilities found in parts (*a*) and (*b*) in terms of percentages.

Table P2.14

		Rank				
		R_1	R_2	R_3	R_4	Total
Age	Under 30 A_1	5	8	57	7	77
	30–39 A_2	50	170	160	20	400
	40–49 A_3	150	120	60	6	336
	50–59 A_4	150	60	30	5	245
	60 & over A_5	70	20	5	0	95
	Total	425	378	312	77	1192

P2.15 Refer to Problem P2.14. Compute the conditional probability $P(A_4/R_3)$ using unconditional probabilities.

P2.16 Data on the marital status of employees in a company are as given in the table below. The table provides a joint probability distribution for the marital status of adults by sex. The term "Single" is used for "Never Married".

Table P2.16

		Marital status				
		Single M_1	Married M_2	Widowed M_3	Divorced M_4	$P(S_1)$
Sex	Male S_1	0.10	0.30	0.01	0.04	0.45
	Female S_2	0.15	0.35	0.03	0.02	0.55
	$P(M_1)$	0.25	0.65	0.04	0.06	1.00

If an employee is selected at random,

(*a*) find the probability that the employee selected is divorced, given that the employee selected is male, given that the employee selected is divorced.

(*b*) use the conditional probability rule for (*a*).

P2.17 In the research and development department of a company, the number of male and female employees are shown in the frequency distribution as given below in a tabular form.

Sex	Frequency
Male	30
Female	20
Total	50

Two employees selected at random from this department. The first employee obtained is not returned to the department for possible reselection; that is, the sampling is without replacement.

(*a*) determine the probability that the first employee obtained is female and the second is male

(*b*) draw the tree diagram.

P2.18 A bin contains 20 machine parts, 4 of which are defective. If 2 machine parts are selected at random (without replacement) from this box, what is the probability that both parts are defective? Draw the tree diagram.

P2.19 Table P2.19 gives the classification of all employees of a particular department in a large corporation by gender and college degree.

Table P2.19

	College graduate (G)	Not a college graduate (N)	Total
Male (M)	8	20	28
Female (F)	3	9	12
Total	11	29	40

If one of these employees is selected at random for membership on the planning committee, determine the probability that this employee is a female and a college graduate. Draw the tree diagram.

P2.20 A balanced coin is tossed five times. Determine the probability of getting at least one head.

P2.21 The probability that a patient is allergic to penicillin is 0.10. If the drug is administered to three individual patients, determine the probability that

(a) all three of them are allergic to it

(b) at least one of them is not allergic to it.

(c) draw the tree diagram.

P2.22 Draw the Venn and tree diagram for selecting three times two persons (a man or a woman) from the members of a club.

P2.23 The search committee of the school of engineering that plans to hire one new faculty member has prepared a final of five candidates, all of whom are equally qualified. Out of these five candidates two of them are woman. If the committee decides to select at random one person out of these five candidates, find the probability that this person will be a woman or a man? Also find the sum of these two probabilities and comment the result.

P2.24 In a random sample of 1000 companies surveyed showed that 35 of them have no design engineer, 210 have one design engineer each, 379 have two design engineers, 270 have three design engineers and 106 have four or more design engineers.

(a) prepare a frequency distribution table and calculate the relative frequencies for all categories

(b) if one company is selected randomly from these 1000 companies, find the probability that this company has to design engineers

(c) find the probability that the company selected has four or more design engineers.

P2.25 Of the 25 new graduates hired by a company, 14 are female and 9 are mechanical engineering majors. Of the 14 females, 6 are mechanical engineering majors.

(a) are these events "female" and "mechanical engineering major" independent?

(b) are these events "mutually exclusive"?

P2.26 An individual is picked at random from a group of 56 athletes. Suppose 26 of the athletes are female and 6 of them are swimmers. Also, there are 10 swimmers among the males.

(*a*) find the probability that she is a swimmer if that the individual picked is a female

(*b*) given that the individual picked is a swimmer, find the probability that the person is a male.

P2.27 (*a*) The probability that the stock market goes up on Monday is 0.7. Given that it goes up on Monday, the probability that it goes up on Tuesday is 0.3. Find the probability that the market goes up on both days.

(*b*) The probability that a student passes the midterm exam is 0.7. Given that a student fails the midterm exam, the probability that the student passes the final exam is 0.8. Find the probability that John (a student in that class) fails the midterm and passes the final.

(*c*) Past records show that the probability that the Chairman of the board attends a meeting is 0.7, that the President of the company attends a meeting is 0.8, and that they both attend a meeting is 0.6. Can we say that the Chairman and the President act independently regarding their attendance at the meeting of the board of directors?

P2.28 Two students are selected at random from a class without replacement whose frequency distribution of males and females is given in Table P2.28.

Table P2.28

Gender	Frequency
Male	17
Female	23
Total	40

(*a*) find the probability that the first student selected is female and the second is male

(*b*) draw a tree diagram.

P2.29 Given that A and B are two independent events, find their joint probability for the following data:

(*a*) $P(A) = 0.25$ and $P(B) = 0.78$

(*b*) $P(A) = 0.55$ and $P(B) = 0.23$

P2.30 The following data gives a two-way classification of all faculty members of a university based on gender and tenure.

	Tenured	Non-tenured
Male	75	48
Female	34	19

(*a*) if one of these faculty members is selected at random, find the following probabilities:

- (*i*) P(male and non-tenured)
- (*ii*) P(tenured and female)

(*b*) find P(tenured and non-tenured).

P2.31 In a group of 12 employees of a company, 5 are graduates and 7 are non-graduates. If 2 persons are selected at random from this group, find the probability that the first of them is a graduate and the second is a non-graduate. Draw a tree diagram for the solution.

P2.32 The probability that an employee at a university is a female is 0.3. The probability that an employee is a female and married is 0.2. Find the conditional probability that a randomly selected female employee from this university is a married person.

P2.33 For the data given in Problem P2.30, if one of these faculty members is selected at random, find the following probabilities.

(a) P(female or non-tenured)

(b) P(tenured or male).

P2.34 The probability that a new product introduced to the market is successful is 0.78. What is the probability that in two randomly selected new products introduced, at least one will be successful? Draw a tree diagram for the solution.

P2.35 Consider the following events for one roll of a die

A = an odd number is observed (1, 3, 5)

B = an even number is observed (2, 4, 6)

C = a number less than 5 is observed (1, 2, 3, 4)

(a) Are events A and B mutually exclusive?

(b) Are events B and C mutually exclusive?

P2.36 An experimental submarine is estimated to have a probability of 0.90 for a successful first flight. In case of failure there is a 0.03 probability of a catastrophic explosion, in which case the abort system cannot be used. The abort system has a reliability of 0.97. Calculate the probability of every possible outcome of the initial flight.

P2.37 With voltage fluctuations present, the probability of an electromechanical system going down is 75%; the probability of it going down during no voltage fluctuations condition is only 8%. The probability of voltage fluctuations occurring is 20%. What is the probability that voltage fluctuations are present given that the electro mechanical system is down?

P2.38 Out of a collection of 100 gaskets the following are defective as indicated:

Defective type	Number of gaskets
Type I defect	10
Type II defect	8
Type III defect	3
Type I and II defect	5
Type I, II and III defect	4
No defect	70

(a) if a gasket picked out the collection has a type II defect, what is the probability that it also has a type I defect?

(b) if the gasket picked has a type I defect, what is the probability that it also has a type II defect?

P2.39 Three persons: A, B and C are working independently to solve a mathematical puzzle. Based on their previous performance, the probabilities that they will succeed are 0.1, 0.2 and 0.3 respectively. What is the probability that the puzzle will be solved?

P2.40 A system consists of 25 different components, of which one; component A, has been identified as the critical component. The probability that the critical component is non-defective is 0.97. If the critical component is non-defective, then the system succeeds with at probability of 0.99. If not, then the probability of the system success drops to 0.93. What is the probability of system success?

P2.41 An electronic test set up has 97% probability of correctly identifying a defective component and a 3% probability of identifying a non-defective component as defective. A batch of 100 components of which 3 are known as defective is subjected to this electronic testing. If the test identifies a component as defective, what is the probability that it is truly defective?

P2.42 A mechanical system assembly plant purchases a certain component from three independent vendors: *A*, *B* and *C*, who supply 20%, 30% and 50% respectively of the total number of components needed by the assembly plant. On an average, the percentage of defective components supplied by vendors *A*, *B* and *C* are 5%, 3% and 2% respectively.

(*a*) if a component is selected at random, find the probability that it is defective

(*b*) if the selected component is defective, what is the probability that it was supplied by vendor *C*? by vendor *A*? by vendor *B*?

P2.43 A bowl contains five white balls, two red balls, and three green balls. What is the probability of getting either a white ball or a red ball in one draw from the bowl?

P2.44 One bowl contains five white balls, two red balls, and three green balls. Another bowl contains three yellow balls and seven black balls. Determine the probability of getting a red ball from the first bowl and a yellow ball from the second bowl in one draw from each bowl.

P2.45 The probability that a student is in favour of engineering ethics course is 0.60 and a student is against is 0.40. Two students are randomly selected, and it is observed whether they favour or oppose engineering ethics course.

(*a*) draw a tree diagram for this experiment

(*b*) determine the probability that at least one of the two students favours engineering ethics course.

P2.46 Customers who buy a certain make of an automobile can order an engine in any of three sizes. Of all automobiles sold, 50% have the smallest engine, 30% have the medium-size one, and 20% have the largest. Of automobiles with the smallest engine, 15% fail the emissions test within two years of purchase, while 10% of those with the medium size and 15% of those with the largest engine fail. Determine the probability that a randomly chosen automobile will fail on emissions test within two years.

P2.47 A company manufactures aluminium soft drink cans. It was found that the probability that a can has a flaw on its side is 0.03, the probability that a can has a flaw on the top is 0.02, and the probability that a can has a flaw on both sides and the top is 0.01.

(*a*) what is the probability that a randomly chosen can has a flaw?

(*b*) what is the probability that it has no flaw?

(*c*) what is the probability that a can will have a flaw on the side, given that it has a flaw on top?

(*d*) what is the probability that a can will have a flaw on the top, given that it has a flaw on the side?

P2.48 (*a*) Two coins are tossed. Determine the conditional probability of getting two heads (event *B*) given that at least one coin shows a head (event *A*)

(*b*) A box contains 5 red and 4 blue balls. Two balls are drawn one-by-one without replacement. Given that the first ball drawn is red, find the probability that both the balls drawn will be blue.

P2.49 Of the gas turbine engine blades manufactured by a certain process, 15% are defective. Five turbine blades are chosen at random. Assume they function independently. Find the probability that they all work.

P2.50 Refer Problem P2.49. What is the probability that at least one of the gas turbine engine blade manufactured works?

P2.51 In a certain city, 45% of the people consider themselves conservatives (C), 30% Consider themselves to be Liberals (L) and 25% consider themselves to be Independents (I). During the particular election, 50% of the Conservatives voted, 40% of the Liberals voted, and 60% of the Independents voted. If a randomly selected person voted, determine the probability that the voter is

(a) Conservative

(b) Liberal

(c) Independent.

P2.52 In a certain company employees, 5% of the men and 1% of the women are heavier than 100 kg. Furthermore, 70% of the employees are women. Suppose a randomly selected employee is heavier than 100 kg. Find the probability that the employee is a woman.

P2.53 In a certain engineering college hostel, 20% are freshmen of whom 10% own a car, 30% are sophomores of whom 20% own a car, 20% are juniors of whom 30% own a car, and 20% are seniors of whom 40% own a car. A student is randomly selected from the hostel.

(a) find the probability that the student owns a car

(b) if the student owns a car, find the probability that the student is a junior.

P2.54 At a certain grocery store, eggs come in cartons that hold a dozen eggs. The store experience indicates that 78% of the cartons have no broken eggs, 19% have one broken egg, 2% have two broken eggs, and 1% have three broken eggs, and that the percentage of cartons with four or more eggs broken eggs is negligible. An egg selected at random from a carton is found to be broken. What is the probability that this egg is the only broken one in the carton?

P2.55 The probability that a college student being male is 1/3 and that being female is 2/3. The probability that a male student completes the probability and statistics course successfully is 8/10 and that the female student does it is 4/5. A student is selected at random is found to have completed the course. Determine the probability that the student is a

(a) male

(b) female.

P2.56 In a car dealership, assume 55% of the cars are made in US and 15% of these are compact; 25% of the cars are made in Europe and 40% of these are compact; and finally, 20% are made in Japan and 60% of these are compact. If a car is picked at random from this dealership:

(a) find the probability that it is a compact

(b) draw a tree diagram for part *(a)*

(c) given that the car is a compact, find the probability that it is a European.

P2.57 In a survey of the evaluation of preliminary product designs by the customers, 90% of highly successful products received good reviews, 60% moderately successful products received good reviews, and 10% of poor products received good reviews. In addition, 40% of products have been highly successful, 35% have been moderately successful, and 25% have been poor products.

(a) find the probability that a product attains a good review

(b) if the new design attains a good review, what is the probability that it will be a highly successful product?

(c) if a product does not attain a good review, what is the probability that it will be a highly successful product?

P2.58 A cat is in a room and each of the four walls of the room has a door through which the cat could try to escape. However, there is a trap at each of the doors D_1, D_2, D_3 and D_4 and they work with probabilities 0.3, 0.2, 0.4 and 0.5 respectively. If the cat picks a door at random:

(a) find the probability that the cat will escape

(b) draw the tree diagram for the solution in part *(a)*

(c) given that the cat escapes, find the probability that the cat chose door D_3 to escape.

P2.59 According to a particular survey, 8% of the population has a lung disease. Of those having lung disease, 90% are smokers; of those not having lung disease, 25% are smokers. Find the probability that a randomly selected smoker has lung disease.

P2.60 Three different machines m_1, m_2 and m_3 were used for producing a large batch of similar manufactured items. Suppose that 30% of the items were produced by machine m_1, 30% by machine m_2, and 40% by machine m_3. Also, 1% of the items produced by machine m_1 are defective, 2% of the items produced by machine m_2 are defective, and 3% of the items produced by machine m_3 are defective. If one item is selected at random from the entire batch, and it is found to be defective. Determine the probability that this item was produced machine m_2.

REVIEW QUESTIONS

1. Define the following terms:
   (a) experiment
   (b) outcome
   (c) sample space
   (d) simple event
   (e) compound event.
2. Describe the properties of probability.
3. Describe an impossible event and a sure event. What is the probability of the occurrence of each of these two events?
4. Describe the three approaches to probability.
5. Explain the difference between the marginal and conditional probability of events.
6. What is meant by two mutually exclusive events?
7. Explain the meaning of independent and dependent events.
8. What is the complement of an event? What is the sum of the probabilities of two complementary events?
9. Explain the meaning of the following terms:
   (a) intersection of two events
   (b) the joint probability of two or more events
   (c) multiplication rule of probability
   (d) joint probability of two mutually exclusive events.
10. Explain the meaning of the following:
    (a) union of two events
    (b) addition rule of probability
    (c) classical probability rule
    (d) equally likely outcomes
    (e) sample point.

STATE TRUE OR FALSE

1. The probability of an event is always between 0 and 1, inclusive. (True/False)
2. The probability of an event that cannot occur is 1. (True/False)
3. The probability of an event that must occur is 1. (True/False)
4. An event that cannot occur is called a certain event. (True/False)
5. An event that must occur is called a certain event. (True/False)
6. An experiment is an action whose outcome cannot be predicted with certainty. (True/False)
7. An event is some specified result that may or may not occur when an experiment is performed. (True/False)
8. The experiment has a finite number of possible outcomes, all equally likely. (True/False)
9. The probability of an event equals the ratio of the number of ways that the event can occur to the total number possible outcomes. (True/False)
10. If a member is selected at random from a finite population, probabilities are identical to percentages (relative frequencies). (True/False)
11. The probability of an event is the proportion of times it occurs in a small number of repetitions of the experiment. (True/False)
12. Sample space is the collection of all possible outcomes for an experiment. (True/False)
13. An event is a collection of outcomes for the experiment, that is, any subset of the sample space. (True/False)
14. Two or more events are mutually exclusive events if no three of them have outcomes in common. (True/False)
15. If event A and event B are mutually exclusive, so are events A, B and C for every event C. (True/False)
16. If event A and event B are not mutually exclusive, neither are events A, B and C for every event C. (True/False)
17. If E is an event, then $P(E)$ represents the probability that event E occurs. (True/False)
18. If event A and event B are mutually exclusive, then $P(A \text{ or } B) = P(A) + P(B)$. (True/False)
19. For any event E, $P(E) = 1 - P(\text{not } E)$. (True/False)
20. If A and B are any two events, then $P(A \text{ or } B) = P(A) + P(B) - P(A \,\&\, B)$. (True/False)
21. The probability that event B occurs given that event A occurs is called a conditional probability. (True/False)
22. If A and B are any two events with $P(A) > 0$, then $P(B|A) = P(A \,\&\, B)/P(B)$. (True/False)
23. If A and B are any two events, then $P(A \,\&\, B) = P(A)/P(B|A)$. (True/False)
24. Event B is said to be independent of event A if $P(B|A) = P(B)$. (True/False)
25. If A and B are independent events, then $P(A \,\&\, B) = P(A) \cdot P(B)$. (True/False)
26. Two or more events are said to be mutually exclusive if at most one of them can occur when the experiment is performed, that is, if no two of them have outcomes in common. (True/False)

27. For any two events, the probability that one or the other of the events occurs equals the sum of the two individual probabilities. (True/False)
28. For any event, the probability that it occurs equals 1 minus the probability that it does not occur. (True/False)
29. Data obtained by observing values of one variable of a population are called univariate data. (True/False)
30. Data obtained by observing values of two variables of a population are called bivariate data. (True/False)
31. The joint probability equals the product of the marginal probabilities. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **False** 5. **True** 6. **True** 7. **True** 8. **True** 9. **True** 10. **True**
11. **False** 12. **True** 13. **True** 14. **False** 15. **False** 16. **True** 17. **True** 18. **True** 19. **True** 20. **True**
21. **True** 22. **False** 23. **False** 24. **True** 25. **True** 26. **True** 27. **False** 28. **True** 29. **True** 30. **True**
31. **True**

CHAPTER 3

Random Variables and Probability Distributions

The topic of random variables is fundamental to probability and statistics. These concepts are natural extensions of the ideas of variables and relative frequency distributions.

This chapter introduces the basic concepts and definitions of random variables (both discrete and continuous) and their mean and standard deviation, and probability distributions. Permutations and combinations are defined briefly. Discrete probability distributions (hypergeometric, binomial and Poisson distributions) and their mean and standard deviations are derived. Continuous probability distribution, namely, the normal distribution, its properties, mean and variance and the standard normal distribution are presented. Approximating probability is also presented. They include Binomial approximation to the hypergeometric, Poisson approximate to the binomial and normal approximation to the binomial and rule for distribution are presented.

3.1 RANDOM VARIABLES

This section introduces the important concept of a probability distribution, which gives the probability for each value of a variable that is determined by chance. This section also introduces procedures for finding the mean and standard deviation for a probability distribution. A *random variable* is a variable (typically represented by x) that has a single numerical value, determined by chance, for each outcome of a procedure. In other words, a *random variable* is a quantitative variable whose value depends on chance. In Chapter 1, we defined a variable as a characteristic that varies from one member of a population to another. When one or more members are selected at random from the population, the variable, in that context, is called a *random variable*. Therefore, mathematically, a random variable is a function defined on the outcome of the sample space.

A *probability distribution* is a description that gives the probability for each value of the random variable. It is often expressed in the format of a graph, table, or formula. A random variable is classified a discrete or continuous, depending upon the range of its values. (The range of a random variable is the set of values it can assume.)

A *discrete random variable* has either a finite number of values or a countable number of values, where *countable* refers to the fact that there might be infinitely many values, but they can be associated with a counting process. Thus, we often say that a discrete random variable can assume at most a *countable infinite* number of values.

The number of eggs that a hen lays in a day, the salary of a new college graduate, the number of bacteria in a culture, the count of the number of students present in statistics class on a given day, the number of cavities that a child has, the price of gold on the exchange on a particular day, toss of a coin, the number of cars sold at a car dealership in a given month, color of a ball drawn from a collection of balls, etc., these are all examples of discrete random variables.

A *continuous random variable* has infinitely many values, and those values can be associated with measurements on a continuous scale without gaps or interruptions.

The maximum daily temperature, the amount of milk a cow produces in one day, the life of an electric bulb, the measure of voltage for a particular smoke detector battery, the distance by which a sharp shooter misses a target, the height of a person, the weight of a person, and the length of time a person has to wait at a bank counter, diameter of a manufactured shaft, time to failure of a machine component, repair time, the price of an automobile, the time taken to complete a medical examination, duration of a snow storm, etc. provide examples of continuous random variables.

3.1.1 Discrete Random Variables

Recall from Chapter 1 that the relative-frequency distribution or relative-frequency histogram of a discrete variable gives the possible values of the variable and the proportion of times each value occurs. We can extend the actions of relative-frequency distribution and relative-frequency histogram-concepts applying to variables of finite populations – to any discrete random variable. Here, we use the terms *probability distribution* and *probability histogram.*

The probability distribution of a discrete random variable, assuming a finite number of values, can be described by listing all the values that the random variable can assume, together with the corresponding probabilities. Such a listing is called the *probability function* of the random variable.

Probability histogram is a graph of the probability distribution that displays the possible values of a discrete random variable on the horizontal axis and the probabilities of those values on the vertical axis. The probability of each value is represented by a vertical bar whose height equals the probability.

In general, if a random variable X assumes the values $x_1, x_2, \ldots, x_k$, then we can represent the probability that X takes the values x_i by p_i. That is

$$P(X = x_i) = p\,(x_i)$$

The probability function can then be summarised in the form of a table, as shown in Table 3.1. By definition, a probability function gives the probabilities with which the values are assumed by the random variable. Hence, $0 \le p\,(x_i) \le 1$.

Table 3.1: The probability function of a random variable

Value x	Probability p(x)
x_1	$p(x_1)$
x_2	$p(x_2)$
$\vdots$	$\vdots$
x_k	$p(x_k)$
Sum	$\Sigma p(x_i) = 1$

Similarly, a random variable has to assume one of its possible values. Therefore, $\Sigma p(x_i) = 1$.

Properties of a Probability Function:

1. The probability that a random variable assumes a value x_i is always between 0 and 1.

   Hence, $0 \leq p(x_i) \leq 1$ (3.1)

2. The sum of all probabilities $p(x_i)$ is equal to 1.

   Hence, $\Sigma p(x_i) = 1$ (3.2)

Let X be a random variable. If the number of possible values of X is finite or countably infinite, X is called a *discrete random variable*. Each possible outcome is x_i. A number, $p(x_i) = p(X = x_i)$, gives the probability that the random variable X equals a value, x_i. The numbers, $p(x_i)$, for $i = 1, 2, \ldots$, must satisfy the following conditions:

1. $p(x_i) \geq 0$ for all i
2. $\sum_{i=1}^{\infty} p(x_i) = 1$ (3.3)
3. $0 \leq p(x_i) \leq 1$

The collection of pairs $[x_i, p(x_i)]$, $i = 1, 2, \ldots$ is called the *probability distribution* of X and $p(x_i)$ is called the *probability mass function* (or pmf) of X. The probability distribution for the variable X is characterised in Fig. 3.1. The *cumulative distribution function* of a discrete random variable is defined as

$$p(X = x_i) = \sum_{x_i \leq X} p(x_i) \tag{3.4}$$

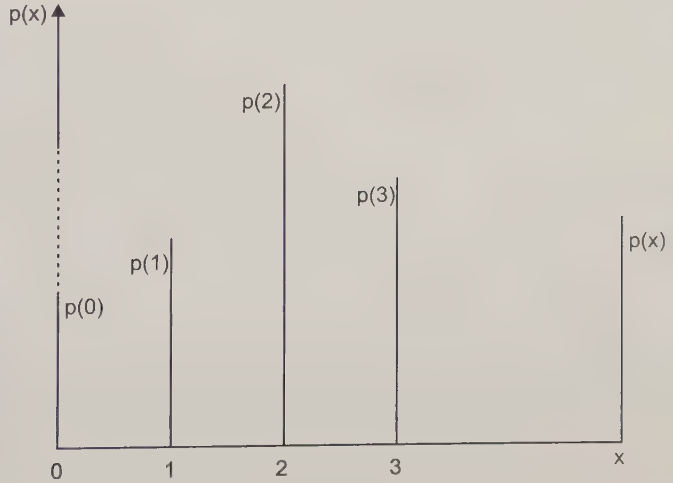


Fig. 3.1: Probability distribution of a discrete random variable x_i

The function is a step function that is constant over every interval not containing any of the points x_i, as shown in Fig. 3.2.

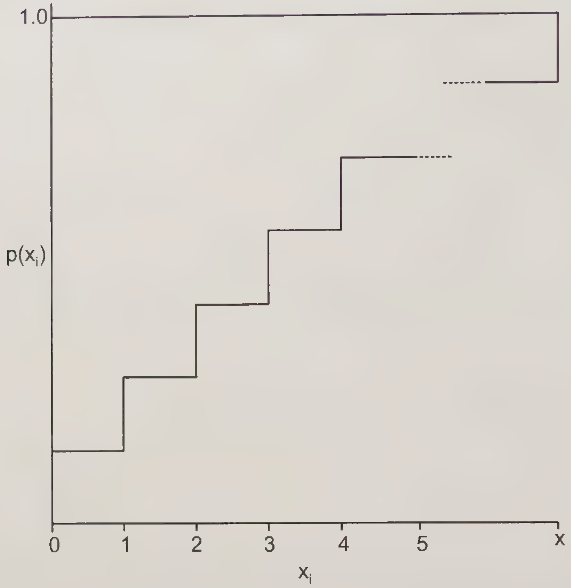


Fig. 3.2: Cumulative distribution function (CDF) of a discrete random variable x_i

3.1.2 Mean and Standard Deviation of a Discrete Random Variable

Consider a set of values, say $[x_i]$, $i = 1, 2, \ldots, n$, each of which occurs with frequency f_i for $i = 1, 2, \ldots, k$. The mean value is the sum of all the x's, hence $\sum_{i=1}^{k} x_i \cdot f_i$ divided by the total number of x's, $\sum_{i=1}^{k} f_i = n$, or

$$\frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i} = \sum_{i=1}^{k} \frac{x_i \cdot f_i}{n} \tag{3.5}$$

But note that this could be written as

$$\frac{\sum_{i=1}^{k} x_i \cdot f_i}{n} = \sum_{i=1}^{k} x_i \cdot \frac{f_i}{n} \tag{3.6}$$

The relative frequencies f_i/n could be regarded as probabilities or values of $f(x)$. Hence,

If a random variable X has the discrete probability function $f(x)$, we define its *mean* or *expected value* or *expectation* as

$$E(X) = \mu = \Sigma x \cdot f(x) \text{ or } \Sigma x \cdot p(x) \tag{3.7}$$

where the summation extends over all the values of x and $p(x)$ are the probabilities.

The *variance* of a random variable X with probability distribution $f(x)$ is given by

$$\sigma^2 = \Sigma(x - \mu)^2 \cdot f(x) \tag{3.8}$$

where the summation extends over all the values of x.

The positive square root of σ^2, σ, is called the *standard deviation.*

We often denote the variance of X by $V(X)$.

It is often easier to use the following formula for σ^2:

$$\sigma^2 = \sum_x x^2 f(x) - \mu^2 \text{ or } \sum_x x^2 p(x) - \mu^2 \tag{3.9}$$

The variance of a random variable X is a measure of the dispersion or scatter in the possible values for X. The variance of X, denoted as σ^2 or $V(X)$ is

$$\sigma^2 = V(X) = \sum_x (x-\mu)^2 f(x) \tag{3.10}$$

$V(X)$ uses weight $f(x)$ as the multiplier of each possible squared deviation $(x - \mu)^2$.

Properties of summations and the definition of μ can be used to show that

$$V(X) = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - 2\mu \sum_x x f(x) + \mu^2 \sum_x f(x)$$

$$= \sum_x x^2 f(x) - 2\mu^2 + \mu^2 = \sum_x x^2 f(x) - \mu^2 \tag{3.11}$$

Therefore, an alternative formula for $V(X)$ can be used.

Summarising, we have

The *mean* or *expected value* of the discrete random variable X, denoted as μ or $E(x)$ is

$$\mu = E(X) = \sum_x x f(x) \tag{3.12}$$

The *variance* of X, denoted as σ^2 or $V(X)$ is

$$\sigma^2 = V(X) = E(X-\mu)^2 = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 f(x) - \mu^2 \tag{3.13}$$

The *standard deviation* of X is

$$\sigma = \sqrt{V(X)} \tag{3.13a}$$

Example E3.1

The number of defects in a gas turbine blade manufactured by a certain process varies from blade to blade. Overall, 50% of the blades produced have no defects, 40% have 1 defect, 8% have 2 defects and 2% have 3 defects. Let x be the number of defects in a randomly selected turbine blade manufactured by the process.

(*a*) find $p(x = 0)$, $p(x = 1)$, $p(x = 2)$ and $p(x = 3)$

(*b*) plot the probability mass function of the random variable x

(*c*) find $F(2)$ and $F(1.5)$

(*d*) plot the cumulative distribution function $F(x)$ of the random variable x that represent the number of defects in a randomly chosen turbine blade

(*e*) population mean of the sample random variable, x

SOLUTION:

(a) $P(x = 0) = 0.50$, $P(x = 1) = 0.40$, $P(x = 2) = 0.08$ and $P(x = 3) = 0.02$

(b) The plot of the probability mass function of the random variable, x is shown in Fig. E3.1(*a*).

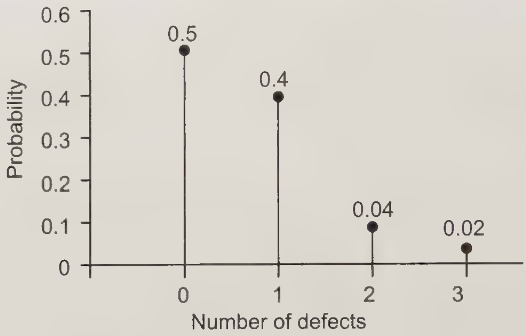


Fig. E3.1(a)

(c) Since $F(2) = P(x \leq 2)$, we need to find $P(x \leq 2)$. We can do this by summing the probabilities for the value of x that are less than or equal to 2, namely 0, 1, and 2. Hence

$$F(2) = P(x \leq 2) = P(x = 0) + P(x = 1) + (x = 2) = 0.50 + 0.40 + 0.08 = 0.98$$

Now $F(1.5) = P(x \leq 1.5)$. Hence, to compute $F(1.5)$, we must sum the probability for the values of x that are less than or equal to 1.5, which are 0 and 1. Therefore,

$$F(1.5) = P(x \leq 1.5) = p(x = 0) + P(x = 1) = 0.50 + 0.40 = 0.90$$

(d) First, we find $F(x)$ for each of the possible values of x, which are 0, 1, 2, and 3.

$F(0) = P(x \leq 0) = 0.50$

$F(1) = P(x \leq 1) = 0.50 + 0.40 = 0.90$

$F(2) = P(x \leq 2) = 0.50 + 0.40 + 0.08 = 0.98$

$F(3) = P(x \leq 3) = 0.50 + 0.40 + 0.08 + 0.02 = 1$

The plot of $F(x)$ is presented in Fig. E3.1(*b*).

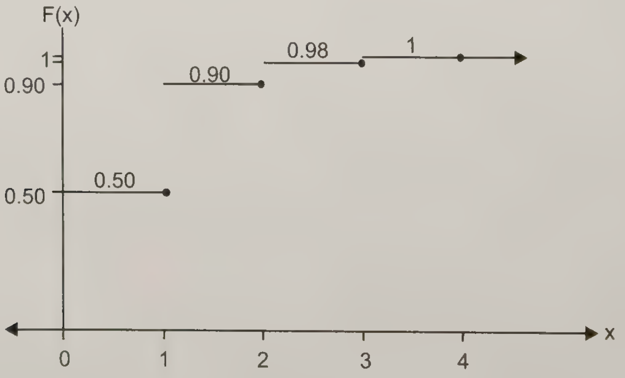


Fig. E3.1(b)

(e) The sample mean is the total number of defects divided by 100.

$$\text{mean} = \frac{0(50)+1(40)+2(8)+3(2)}{100} = 0.62$$

This can be written as

$$\text{mean} = 0(0.50) + 1(0.40) + 2(0.08) + 3(0.02) = 0.62$$

The above calculation for the mean shows that the mean of a perfect sample can be obtained by multiplying each possible value of x by its probability, and summing the products, which is the definition of the population mean of a discrete random variable.

Example E3.2

Table E3.2 gives the probability distribution of a discrete random variable x.

Table E3.2

x	0	1	2	3	4	5
P(x)	0.03	0.13	0.22	0.30	0.20	0.12

Find the following probabilities:

(a) $P(x=1)$, $P(x \le 1)$, $P(x \le 3)$ and $P(0 \le x \le 2)$

(b) probability that x assumes a value less than 3

(c) probability that x assumes a value greater than 3

(d) probability that x assumes a value in the interval 2 and 4.

SOLUTION:

(a) $P(x=1) = 0.13$

$P(x \le 1) = P(0) + P(1) = 0.03 + 0.13 = 0.16$

$P(x \le 3) = P(3) + P(4) + P(5) = 0.30 + 0.20 + 0.12 = 0.62$

$P(0 \le x \ge 2) = P(0) + P(1) + P(2) = 0.03 + 0.13 + 0.22 = 0.38$

(b) $P(x<3) = P(0) + P(1) + P(2) = 0.03 + 0.13 + 0.22 = 0.38$

(c) $P(x>3) = P(4) + P(5) = 0.20 + 0.12 = 0.32$

(d) $P(2 \le x \le 4) = P(2) + P(3) + P(4) = 0.22 + 0.30 + 0.20 = 0.72$

Example E3.3

Table E3.3 lists the probability distribution of x, where x is the number of car accidents that occur in a city during a week.

Table E3.3

x	0	1	2	3	4	5	6
P(x)	0.12	0.16	0.22	0.17	0.15	0.12	0.06

(a) draw a line graph for the probability distribution

(*b*) find the probability that the number of car accidents that will occur during a given week in the city is
(*i*) exactly 4 (*ii*) at least 3 (*iii*) less than 3 (*iv*) 3 to 5

SOLUTION:

(*a*) The line graph is shown in Fig. E3.3.

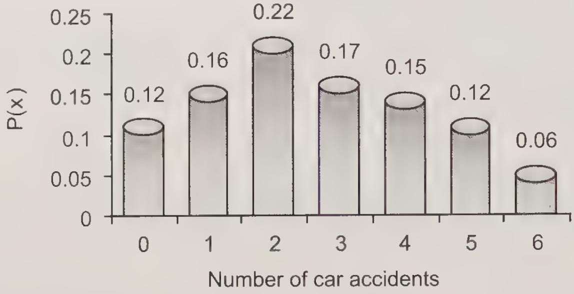


Fig. E3.3

(*b*) (*i*) $P(\text{exactly } 4) = P(4) = 0.15$

(*ii*) $P(\text{at least } 3) = P(x \geq 3) = P(3) + P(4) + P(5) + P(6) = 0.17 + 0.15 + 0.12 + 0.06 = 0.50$

(*iii*) $P(\text{less than } 3) = P(x < 3) = P(0) + P(1) + P(2) = 0.12 + 0.16 + 0.22 = 0.50$

(*iv*) $P(3 \text{ to } 5) = P(3) + P(4) + P(5) = 0.17 + 0.15 + 0.12 = 0.44$

Example E3.4

According to a survey, 75% of all elementary school teachers in a city in the year 2005 were women. Assume that this result holds true for the current population of all teachers in that city. Suppose 2 teachers are randomly selected from the population of all teachers in that city. Denoting x as the number of women in that sample, construct the probability distribution table of x. Draw a tree diagram for this problem.

SOLUTION:

Let M = teacher selected is a man

W = teacher selected is a woman

Then $P(W) = 0.75$ and $P(M) = 1 - P(W) = 1 - 0.75 = 0.25$

The tree diagram is shown in Fig. E3.4.

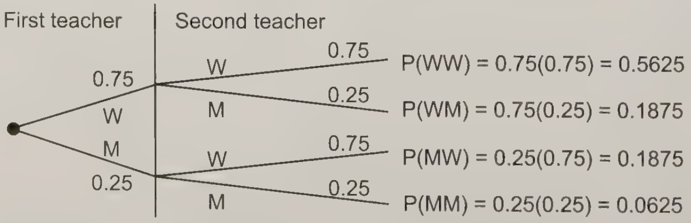

First teacher
Second teacher
0.75
W
M
0.25
W
0.75
M
0.25
W
0.75
M
0.25
P(WW) = 0.75(0.75) = 0.5625
P(WM) = 0.75(0.25) = 0.1875
P(MW) = 0.25(0.75) = 0.1875
P(MM) = 0.25(0.25) = 0.0625

Fig. E3.4

Now, x = the number of women in a sample of 2 teachers. Table E3.4 lists the probability distribution of x, where $x = 0$ represents if neither teacher is a woman, $x = 1$ if the teacher is a woman and one is a man, and

$x = 2$ if the both teachers are woman. The probability in the table are obtained from the tree diagram in Fig. E3.4. The probability of $x = 1$ is obtained by adding the probabilities of *WM* and *MW*.

Table E3.4

Outcomes	x	P(x)
MM	0	0.0625
WM or MW	1	0.3750
WW	2	0.5625

Example E3.5

A machine design class has 20 students and 11 of them are female. Suppose 2 students are randomly selected from the class and x denotes the number of females in the sample.

(*a*) draw a tree diagram

(*b*) find the probability distribution of x.

Assume here that the draws are made without replacement from a small population and the probabilities of outcomes do not remain constant for each draw.

SOLUTION:

Let A = first student selected is a male

B = first student selected is a female

C = second student selected is a male

D = second student selected is a female

(*a*) The tree diagram is shown in Fig. E3.5.

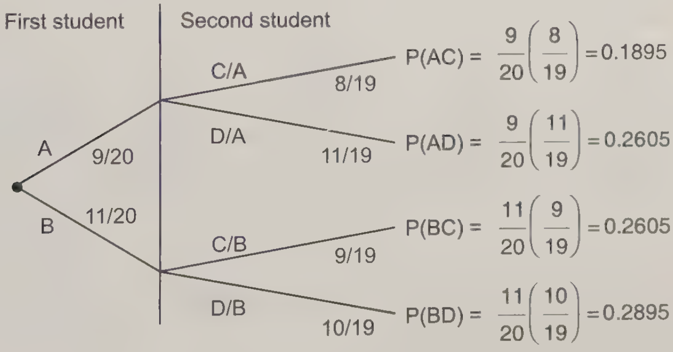

First student
Second student
A
9/20
B
11/20
C/A
8/19
D/A
11/19
C/B
9/19
D/B
10/19
P(AC) = $\frac{9}{20}\left(\frac{8}{19}\right)$ = 0.1895
P(AD) = $\frac{9}{20}\left(\frac{11}{19}\right)$ = 0.2605
P(BC) = $\frac{11}{20}\left(\frac{9}{19}\right)$ = 0.2605
P(BD) = $\frac{11}{20}\left(\frac{10}{19}\right)$ = 0.2895

Fig. E3.5

(*b*) Table E3.5 lists the probability distribution of x, where x is the number of females in a sample of 2 students.

Table E3.5

Outcomes	x	P(x)
AC	0	0.1895
AD or BC	1	0.5210
BD	2	0.2895

Example E3.6

Table E3.6 lists the probability distribution of the number of breakdowns per week for a machine based on past data.

Table E3.6

Breakdowns per week	0	1	2	3
Probability	0.15	0.25	0.40	0.20

(a) present this probability distribution graphically

(b) find the probability that the number of breakdowns for this machine during a given week is

(i) exactly 2 *(ii)* 0 to 2 *(iii)* more than 1 *(iv)* at most 1.

SOLUTION:

Let x denotes the number of breakdowns for this machine during a given week. The Table E3.6 lists the probability distribution of x.

Table E3.6: Probability distribution of breakdowns

x	P(x)
0	0.15
1	0.25
2	0.40
3	0.20
	$\Sigma P(x) = 1.00$

(a) Figure E3.6 shows the line graph of the probability distribution.

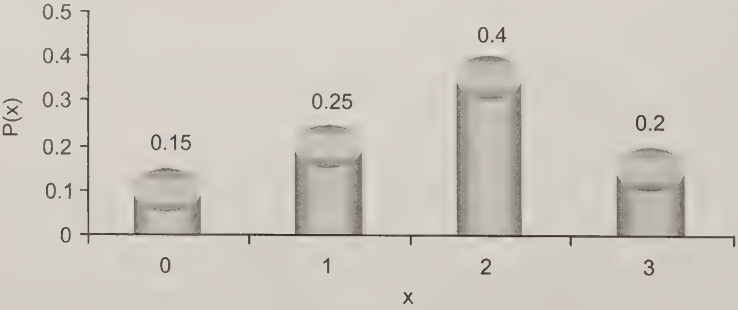


Fig. E3.6: Graphical representation of the probability distribution

(b) *(i)* The probability of exactly, two breakdowns is

$P(\text{exactly 2 breakdowns}) = P(x=2) = 0.40$

(ii) The probability of 0 to 2 breakdowns is given by the sum of the probabilities of 0, 1 and 2 breakdowns

$$P(\text{0 to 2 breakdowns}) = P(0 \le x \le 2)$$
$$= P(x=0) + P(x=1) + P(x=2) = 0.15 + 0.25 + 0.40 = 0.80$$

(iii) The probability of more than 1 breakdown is obtained by adding the probabilities of 2 and 3 breakdowns

$$P(\text{more than 1 breakdown}) = P(x > 1)$$
$$= P(x=2) + P(x=3) = 0.40 + 0.20 = 0.60$$

(*iv*) The probability of at most 1 breakdown is given by the sum of the probabilities of 0 and 1 breakdowns.

$$P(\text{at most 1 breakdown}) = P(x \le 1)$$

$$= P(x = 0) + P(x = 1) = 0.15 + 0.25 = 0.40$$

Example E3.7

Table E3.7 lists the probability distribution of x, where x is the number of defects contained in a randomly selected manufactured product. Find the mean and standard deviation of x.

Table E3.7

x	0	1	2	3	4
P(x)	0.41	0.12	0.05	0.03	0.01

SOLUTION:

Refer to Table E3.7(a)

Table E3.7(a)

x	P(x)	x P(x)	x^2 P(x)
0	0.41	0	0
1	0.12	0.12	0.12
2	0.05	0.10	0.20
3	0.03	0.09	0.27
4	0.01	0.04	0.16
Total		0.35	0.75

Mean $\mu = \Sigma x P(x) = 0.35$ defects (from Table E3.7(a)).

Standard deviation $\sigma = \sqrt{x^2 P(x) - \mu^2} = \sqrt{0.75 - 0.35^2} = 0.79$ defects.

Example E3.8

Find the mean, the variance, and the standard deviation of x, where x denotes the number that shows up when a fair die is rolled.

SOLUTION:

The probability function of x and the necessary calculations are summarised in Table E3.8.

Table E3.8

x	p(x)	x p(x)	x^2p(x)
1	1/6	1/6	1/6
2	1/6	2/6	4/6
3	1/6	3/6	9/6
4	1/6	4/6	16/6
5	1/6	5/6	25/6
6	1/6	6/6	36/6
Σ	1	21/6 = 3.5	91/6 = 15.1667

The mean given by

$$\mu = \Sigma x p(x) = \frac{21}{6} = 3.5$$

Variance is given by

$$\sigma^2 = \Sigma x^2 p(x) - \mu^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = 2.92$$

The standard deviation = $\sqrt{\sigma^2} = \sqrt{2.92} = 1.71$

Hence, mean = 3.5, variance = 2.92 and standard deviation = 1.71.

3.1.3 Continuous Random Variables

A random variable is *continuous* if its probabilities are given by areas under a curve. The curve is called a *probability density function* for the random variable. The probability density function is sometimes called the *probability distribution.*

If the range space of the random variable X is an interval or a collection of intervals, X is called a *continuous random variable*. If a and b are real numbers and $a < b$, then the probability that X lies in the interval $[a, b]$ is defined as

$$P[a \le X \le b] = \int_a^b f(x)\, dx \tag{3.14}$$

The function $f(x)$ is known as the *probability density function* (or pdf) of the random variable X. The pdf must satisfy the following conditions:

1. $f(x) \ge 0, \qquad -\infty < x < \infty$
2. $\int f(x)\, dx = 1$ (3.15)

The probability density function (pdf) is shown in Fig. 3.3. The shaded area in Fig. 3.3 represents the probability that X lies in the interval $[a, b]$.

From Eq. (3.14), for any specified value of x, $P(X = x) = 0$, since

$$\int_x^x f(t)\, dt = 0 \tag{3.16}$$

Hence, we can also write

$$P(a \le X \le b) = P(a < X \le b) = P(a \le X < b) = P(a < X < b) \tag{3.17}$$

Since x is continuous

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{3.18}$$

The following conditions are satisfied

1. F is a non-decreasing function. If $a < b$, then $F(a) \le F(b)$.
2. $\lim\limits_{x \to \infty} F(x) = 1$

3. $\lim_{x \to -\infty} F(x) = 0$ (3.19)

All probability questions about X can be expressed in terms of the cdf. For instance

$$P(a < X \leq b) = F(b) - F(a) \text{ for } a < b \tag{3.20}$$

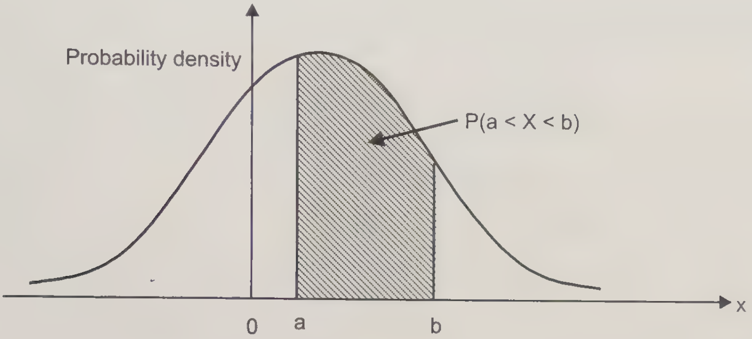


Fig. 3.3: The probability density function (pdf)
(The shaded area under the curve is P(a < X < b))

The cumulative distribution function of a continuous random variable X is $F(X) = P(X \leq x)$, just like for a discrete random variable. For a continuous variable, the value of $F(x)$ is obtained by integrating the probability density function.

Since $F(x) = P(X \leq x)$, we have

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{3.21}$$

where $f(t)$ is the probability density function.

Hence, the continuous distribution function (cdf) of X is the function

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt \tag{3.22}$$

3.1.4 Mean and Variance for Continuous Random Variables

The population mean and variance of a continuous random variable are defined in the same way as those of a discrete random variable, except that the probability density function is used instead of the probability mass function. Specifically, if X is a continuous random variable, its population mean is defined to be the center of mass of its probability density function and its population variance is the moment of inertia around a vertical axis through the population mean.

Let X be a continuous random variable with probability density function $f(x)$. Then, the mean is given by

$$\mu_x = \int_{-\infty}^{\infty} x\, f(x)\, dx \tag{3.23}$$

The *mean* of X is sometimes called the *expectation* or the *expected* value of X and denoted by $E(X)$ or by μ.

The variance of X is given by

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x-\mu_x)^2 f(x)\,dx \tag{3.24}$$

An alternative formula for the variance is given by

$$\sigma_X^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu_X^2 \tag{3.25}$$

The variance of X is also denoted by $V(X)$ or by σ^2.

The standard deviation is the square root of the variance

$$\sigma_X = \sqrt{\sigma_X^2} \tag{3.26}$$

Alternate Formula for the Variance

Starting with the equation

$$\sigma_X^2 = \sum_x (x-\mu_x)^2 P(X=x) \tag{3.27}$$

Expanding the above equation, we obtain

$$\sigma_X^2 = \sum_x (x^2 - 2x\mu_X + \mu_X^2) P(X=x) \tag{3.28}$$

or

$$\sigma_X^2 = \sum_x [x^2 P(X=x) - 2x\mu_X P(X=x) + \mu_X^2 P(X=x)] \tag{3.29}$$

Summing the terms separately, we have

$$\sigma_X^2 = \sum_x x^2 P(X=x) - \sum_x 2x\mu_X P(X=x) + \sum_x \mu_X^2 P(X=x) \tag{3.30}$$

Noting that

$$\sum_x 2x\mu_X P(X=x) = 2\mu_X \sum_x xP(X=x) = 2\mu_X\mu_X = 2\mu_X^2 \tag{3.31}$$

and

$$\sum_x \mu_X^2 P(X=x) = \mu_X^2 \sum_x P(X=x) = \mu_X^2 (1) = \mu_X^2 \tag{3.32}$$

Substituting Eqs. (3.31) and (3.32) in Eq. (3.30), we get

$$\sigma_X^2 = \sum_x x^2 P(X=x) - 2\mu_X^2 + \mu_X^2 \tag{3.33}$$

or

$$\sigma_X^2 = \sum_x x^2 P(X=x) - \mu_X^2 \tag{3.34}$$

3.1.5 Expectation

A brief summary is given here:

If X is a random variable, the expected value of X, denoted by $E(X)$, for discrete and continuous variable is defined as

$$E(X) = \sum_{\text{all } i} x_i\, p(x_i) \qquad \text{if } X \text{ is discrete} \tag{3.35}$$

and
$$E(X) = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \text{if } X \text{ is continuous} \tag{3.36}$$

the expected value $E(X)$ of a random variable X is also referred to as the *mean* of X, μ_X (or μ), or the *first moment* of X. The quantity, $E(X^n)$, $n \geq 1$, is called the n^{th} *moment* of X and is defined as

$$E(X^n) = \sum_{\text{all } i} x_i^n\, p(x_i) \qquad \text{if } X \text{ is discrete} \tag{3.37}$$

and
$$E(X^n) = \int_{-\infty}^{\infty} x^n f(x)\, dx \qquad \text{if } X \text{ is continuous} \tag{3.38}$$

The variance of a random variable, X, denoted by $V(X)$, $\mathrm{Var}(X)$, or σ^2, is defined by

$$V(X) = E([X - E(X)]^2)$$

A useful identity in computing $V(X)$ is given by

$$V(X) = E(X^2) - [E(X)]^2 \tag{3.39}$$

The mean $E(X)$ is a measure of the central tendency of a random variable. The variance, $V(X)$, is a measure of the spread or dispersion of the possible values of X around the mean $E(X)$. The standard deviation, σ, is the square root of the variance.

Example E3.9

The life of an electronic component is given by X, a continuous random variable assuming all values in the range $x \geq 0$. The pdf of the lifetime, in years is given as

$$f(x) = \begin{cases} \dfrac{1}{4} e^{-\frac{x}{4}}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

(*a*) determine the probability that the life of this component is between 2 and 4 years.

(*b*) determine cdf of the components

(*c*) determine the probability that this electronic component will last for 4 years

(*d*) find the mean and variance of the life of the component.

SOLUTION:

(*a*) $P(2 \leq X \leq 4) = \dfrac{1}{4}\displaystyle\int_2^4 e^{-\frac{x}{4}} dx = -e^{-1} + e^{-\frac{1}{2}} = 0.2386$

The probability that the laser will last between 2 and 4 years is obtained also from

$P(2 \leq X \leq 4) = F(4) - F(2) = (1 - e^{-4/4}) - (1 - e^{-2/4}) = e^{-1/2} - e^{-1} = 0.6065 - 0.3679 = 0.2386$

(*b*) The cdf of this component is given by

$$F(x) = \frac{1}{4}\int_{0}^{x} e^{-\frac{x}{4}}dx = 1 - e^{-\frac{x}{4}}$$

(*c*) The probability that the component will last for 4 years or less is given by

$$P(0 \le X \le 4) = F(4) - F(0) = F(4) = 1 - e^{4/4} = 0.6321$$

(*d*) The mean and variance of the life of the electronic component are determined as follows:

$$E(X) = \frac{1}{4}\int_{0}^{\infty} xe^{-\frac{x}{4}}dx = xe^{-\frac{x}{4}}\Big|_{0}^{\infty} + \int_{0}^{\infty} e^{-\frac{x}{4}}dx = 0 + \frac{1}{1/4}e^{-\frac{x}{4}}\Big|_{0}^{\infty} = 4 \text{ years}$$

To compute $V(X)$, first compute $E(X^2)$ as follows:

$$E(X^2) = \frac{1}{4}\int_{0}^{\infty} x^2 e^{-\frac{x}{4}}dx = x^2 e^{-\frac{x}{4}}\Big|_{0}^{\infty} + 8\int_{0}^{\infty} xe^{-\frac{x}{4}}dx = 32$$

Hence, variance is given by

$$V(X) = 32 - 4^2 = 16 \text{ years}$$

Example E3.10

Given $f(x) = e^{-x}$ for $0 < x$. Find the following probabilities:

(*a*) $P(1 < X)$

(*b*) $P(1 < X < 2.5)$

(*c*) $P(X = 3)$

(*d*) $P(X < 4)$

(*e*) $P(3 \le X)$

SOLUTION:

(*a*) $$P(1 < X) = \int_{1}^{\infty} e^{-x}dx = (-e^{-x})\Big|_{1}^{\infty} = e^{-1} = 0.3679$$

(*b*) $$P(1 < X < 2.5) = \int_{1}^{2.5} e^{-x}dx = (-e^{-x})\Big|_{1}^{2.5} = e^{-1} - e^{-2.5} = 0.2858$$

(*c*) $$P(X = 3) = \int_{3}^{3} e^{-x}dx = 0$$

(*d*) $$P(X < 4) = \int_{0}^{4} e^{-x}dx = (-e^{-x})\Big|_{0}^{4} = 1 - e^{-4} = 0.9817$$

(*e*) $P(3 \le X) = \int_3^{\infty} e^{-x}\,dx = (-e^{-x})\Big|_3^{\infty} = e^{-3} = 0.0498$

Example E3.11

The probability density of the continuous random variable X is given by

$$f(x) = \left.\begin{matrix} 1/5 \\ 0 \end{matrix}\right\}; \quad \begin{matrix} \text{for } 2 < x < 7 \\ \text{elsewhere} \end{matrix}$$

Find (*a*) $P(3 < X < 5)$

(*b*) area under the curve.

SOLUTION:

(*a*) $P(3 < X < 5) = \int_3^5 \frac{1}{5}\,dx = \frac{1}{5}\int_3^5 1dx = \frac{1}{5}(5-3) = \frac{2}{5}$

(*b*) Area under the curve $= \int_{-\infty}^{\infty} f(x)\,dx = \int_2^7 \frac{1}{5}\,dx = \frac{1}{5}x\Big|_2^7 = \frac{1}{5}(7-2) = 1$

Example E3.12

Given $f(x) = 1.5x^2$ for $-1 < x < 1$. Find the mean and variance of x.

SOLUTION:

$$E(X) = \int_{-1}^{1} x(1.5x^2)\,dx = \int_{-1}^{1} 1.5x^3 dx = 1.5\frac{x^4}{4}\Bigg|_{-1}^{1} = 0$$

$$V(X) = \int_{-1}^{1} 1.5x^2(x-0)^2\,dx = 1.5\int_{-1}^{1} x^4 dx = 1.5\frac{x^5}{5}\Bigg|_{-1}^{1} = 0.6$$

Example E3.13

The pdf of the random variable X is given by

$$f(x) = \left.\begin{matrix} \frac{a}{\sqrt{x}} \\ 0 \end{matrix}\right\}; \quad \begin{matrix} \text{for } 0 < x < 4 \\ \text{elsewhere} \end{matrix}$$

Find (*a*) the value of a

(*b*) $P\left(X < \frac{1}{4}\right)$ and $P(X > 1)$.

SOLUTION:

(a) $$\text{Area} = 1 = \int_0^4 \frac{a}{\sqrt{x}}dx = a\int_0^4 x^{-1/2}dx = a\frac{x^{1/2}}{1/2}\bigg|_0^4 = 2a{\cdot}2 = 4a$$

Hence, $a = 1/4$.

(b) $$P\left(X < \frac{1}{4}\right) = \int_0^{1/4} \frac{1}{4\sqrt{x}}dx = \frac{1}{4}\int_0^{1/4} x^{-1/2}dx = \frac{1}{4}\frac{\sqrt{x}}{1/2}\bigg|_0^{1/4} = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}$$

$$P(X > 1) = 1 - \int_0^1 \frac{1}{4\sqrt{x}}dx = 1 - \frac{1}{2}\sqrt{x}\bigg|_0^1 = \frac{1}{2}$$

Example E3.14

The thickness of a conductive coating in micrometers has a density function of $f(x) = 600\,x^{-2}$ for $100\ \mu\text{m} < x < 120\ \mu\text{m}$.

(a) find the mean and variance of the coating thickness

(b) if the cost of coating is \$0.25 per micrometer of thickness on each component, find the average cost of the coating per component.

SOLUTION:

(a) $$E(X) = \int_{100}^{120} x\frac{600}{x^2}dx = 600\ln x\bigg|_{100}^{120} = 109.39$$

$$V(X) = \int_{100}^{120} (x - 109.39)^2\frac{600}{x^2}dx = 600\int_{100}^{120} 1 - \frac{2(109.39)}{x} + \frac{(109.39)^2}{x^2}dx$$

$$= 600(x - 218.75\ln x - 109.39^2 x^{-1})\bigg|_{100}^{120} = 33.19$$

(b) Average cost per component = \$0.25(109.39) = \$27.35.

Example E3.15

The distribution function of the random variable X is given by

$$F(x) = \begin{array}{ll} 1-(1+x)e^{-x} & \text{for } x > 0 \\ 0 & \text{for } x \le 0 \end{array}$$

Find *(a)* $P(X \le 2)$

(b) $P(1 < X < 3)$

(c) $P(X > 4)$

SOLUTION:

(a) $P(X \le 2) = F(2) = 1 - 3e^{-2} = 1 - 3(0.1353) = 1 - 0.4074 = 0.5926$

(b) $P(1 < X < 3) = F(3) - F(1) = 1 - 4e^{-3} - 1 + 2e^{-1} = 2e^{-1} - 4e^{-3} = 2(0.3679) - 4(0.0498) = 0.5366$

(c) $P(X > 4) = 1 - F(4) = 5e^{-4} = 5(0.0183) = 0.0915$

Example E3.16

The pdf of a random variable X is given by

$$f(x) = \begin{matrix} \frac{4}{3}(1-x^3) & \text{for } 0 < x < 1 \\ 0 & \text{elsewhere} \end{matrix}$$

Sketch the pdf and find the values of the following:

(*a*) $P\left(X < \frac{1}{2}\right)$

(*b*) $P\left(\frac{1}{4} < X < \frac{3}{4}\right)$

(*c*) $P\left(X > \frac{1}{3}\right)$

SOLUTION:

The plot is shown in Fig. E3.16.

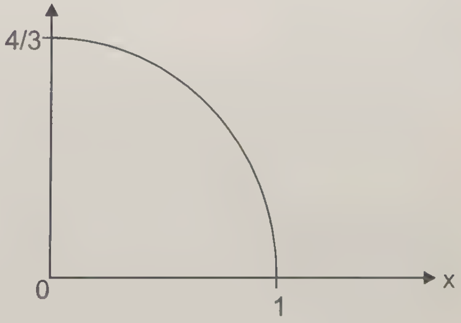


Fig. E3.16 pdf of the function f(x)

(*a*) $P\left(X < \frac{1}{2}\right) = \int_0^{1/2} 4(1-x^3)\frac{dx}{3} = 0.6458$

(*b*) $P\left(\frac{1}{4} < X < \frac{3}{4}\right) = \int_{1/4}^{3/4} 4(1-x^3)\frac{dx}{3} = 0.5625$

(*c*) $P\left(X > \frac{1}{3}\right) = \int_{1/3}^{1} 4(1-x^3)\frac{dx}{3} = 0.5597$

Example E3.17

The pdf of a random variable X is given by

$$f(x) = \begin{cases} ax^2 & \text{for } 0 \le x \le 2 \\ 0 & \text{elsewhere} \end{cases}$$

(*a*) find the value of the constant a and sketch the pdf

(*b*) find $P(X > 3/2)$.

SOLUTION:

(*a*) We must have

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_1^2 a\,x^2\,dx = \frac{7}{3}a = 1$$

Hence, $a = 3/7$. The pdf is shown in Fig. E3.17.

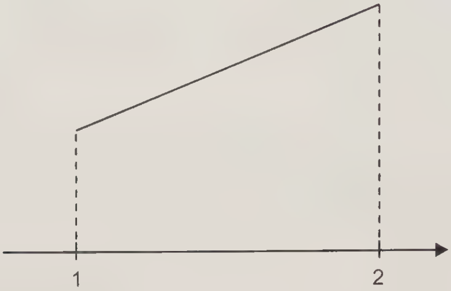


Fig. E3.17

(*b*) $P(X > 3/2) = \int_{3/2}^{2} f(x)\,dx = \int_{3/2}^{2} \frac{3}{7} x^2 dx = 37/56$

3.2 PERMUTATIONS AND COMBINATIONS

Factorial Notation: The product of the positive integers from 1 to n inclusive is denoted by $n!$ (read as "n factorial"). The value of the factorial of a number is obtained by multiplying all integers from that number to 1.

The symbol $n!$, read as "n factorial" represents the product of all integers from n to 1. In other words,

$$n! = n(n-1)(n-2)(n-3) \cdots 3 \cdot 2 \cdot 1 \tag{3.40}$$

It is also convenient to define

$$0! = 1,\ 1! = 1 \text{ and } n! = n \cdot ((n-1)!) \tag{3.41}$$

Example E3.18

Evaluate the following:

(*a*) 2!, 3!, 4!, 5!, 6! and 7!

(*b*) (7 – 3)!, (14 – 12)!, (7 – 2)!, (13 – 4)!, and (6 – 6)!

(*c*) $\frac{8!}{6!}, \frac{12!}{9!}$ and $\frac{12!}{9!3!}$

SOLUTION:

(*a*) $2! = 2 \cdot 1 = 2$

$3! = 3 \cdot 2 \cdot 1 = 6$

$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$

$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$

$6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$

$7! = 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 5040$

(*b*) $(7-3)! = 4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$

$(14-12)! = 2! = 2.1 = 2$

$$(7-2)! = 5! = 5{\cdot}4{\cdot}3{\cdot}2{\cdot}1 = 120$$

$$(13-4)! = 9! = 9{\cdot}8{\cdot}7{\cdot}6{\cdot}5{\cdot}4{\cdot}3{\cdot}.2{\cdot}1 = 362{,}880$$

$$(6-6)! = 0! = 1$$

(c) $$\frac{8!}{6!} = \frac{8{\cdot}7{\cdot}6!}{6!} = 8.7 = 56$$

$$\frac{12!}{9!} = \frac{12{\cdot}11{\cdot}10{\cdot}9!}{9!} = 12{\cdot}11{\cdot}10 = 1320$$

$$\frac{12!}{9!3!} = \frac{12{\cdot}11{\cdot}10{\cdot}9!}{9!3!} = \frac{12{\cdot}11{\cdot}10}{3.2} = 220$$

Example E3.19

Find the value of 9! by using the table in Appendix-*A*.

SOLUTION:

We locate 9 in the column labeled *n*. Then we read the value in the column for *n*! entered next to 9. Hence,

$$9! = 362,\ 880$$

3.2.1 Permutations

Any arrangement of a set of *n* objects in a given order is called a *permutation* of the objects (taken all at a time). Any arrangement of any $r \le n$ of those objects in a given order is called an *r-permutation* or *a permutation of the n objects taken r at a time*.

Ordered arrangement of a set of objects are called *permutations*. When the order in which they are arranged is disregarded, then the arrangements are known as *combinations*. The number of permutations of *n* different items is the number of different arrangements in which these items can be placed. A permutation of *n* different objects taken *r* at a time is an arrangement of *r* out of the *n* objects with attention given to the order of arrangement. The number of permutations of *n* objects taken *r* at a time is denoted by nP_r, $P(n, r)$ P_r^n, $(n)_r$ or $P_{n,r}$ and is written as

$$nP_r = n(n-1)(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!} \tag{3.42}$$

Derivation of the Formula nP_r

The first element in an *r*-permutation of *n* objects can be chosen in *n* different ways. Following this way, the second element in the permutation can be chosen in $n-1$ ways, and the 3rd element in the permutation can be chosen in $n-2$ ways etc. Finally, the r^{th} (last) element in the *r*-permutation can be chosen in $n-(r-1) = n-r+1$ ways. Hence, from the fundamental principle of counting, we have

$$nP_r = P(n, r) = n(n-1)\,(n-2)\cdots(n-r+1) \tag{3.43}$$

or $$n(n-1)\,(n-2)\cdots(n-r+1) = \frac{n(n-1)(n-2)\cdots(n-r+1).(n-r)!}{(n-r)!} = \frac{n!}{(n-r)!} \tag{3.44}$$

in which $n!$ is the symbol for factorial $n = n(n-1)\cdots 3 \times 2 \times 1$, when $r = n$, $n\,P_n = n!$ (note by definition $0! = 1$). Tables of the values of factorial n are given in Appendix-*A*.

The permutation of n items taken all at a time, when the n items consists of r_1 alike, r_2 alike, … r_k alike, so that by definition $r_1 + r_2 + \cdots + r_k = n$.

The number of permutation is then

$$P = \frac{n!}{r_1!r_2!\cdots r_k!} \tag{3.45}$$

Also, we have $nP_n = n(n-1)\cdots 3{\cdot}2{\cdot}1 = n!$

Example E3.20

How many five-digit numbers can be formed from the numbers 1 to 8?

SOLUTION:

Here $n = 8$ and $r = 5$.

$$nP_r = \frac{n!}{(n-r)!} = \frac{8!}{(8-5)!} = \frac{8\times7\times6\times5\times4\times3\times2\times1}{3\times2\times1} = 1680 \text{ numbers.}$$

Example E3.21

How many numbers of permutations of letters are there in the word *quality*?

SOLUTION:

Here $n = 7$ (7 letters in word *quality*)

$r_1 = 1(q)$

$r_2 = 1(u)$

$r_3 = 1(a)$

$r_4 = 1(l)$

$r_5 = 1(i)$

$r_6 = 1(t)$

$r_7 = 1(y)$

Therefore,

$$P = \frac{n!}{r_1!r_2!\cdots r_k!} = \frac{7!}{1!1!1!\cdots 1!} = 7\times6\times5\times4\times3\times2\times1 = 5040$$

Example E3.22

- (*a*) Evaluate $7P_3$, $6P_4$, $16P_1$, $3P_3$ and $6P_3$
- (*b*) five persons arrive at the bank teller window at the same time. In how many different ways can these people line up?
- (*c*) find the number of ways of arranging the letters in the word *subject*
- (*d*) find the number of ways if each arrangement in part (*c*) starts with the letter *j*.

SOLUTION:

(a) $7P_3 = 7{\cdot}6{\cdot}5 = 210$

$6P_4 = 6{\cdot}5{\cdot}4{\cdot}3 = 360$

$16P_1 = 16$

$3P_3 = 3{\cdot}2{\cdot}1 = 6$

$6P_3 = 6{\cdot}5{\cdot}4 = 120$

(b) These paper can line up in different ways

$5! = 5{\cdot}4{\cdot}3{\cdot}2{\cdot}1 = 120$ ways.

(c) There are 7 distinct letters in the word *subject*. Hence, the number of arrangements is 7!. That is

$7! = 7{\cdot}6{\cdot}5{\cdot}4{\cdot}3{\cdot}2{\cdot}1 = 5040$

(d) Since the first letter is fixed as j, the total number of possibilities are found by arranging the remaining six letters in all possible ways. This can be done by 6! ways.

$6! = 6{\cdot}5{\cdot}4{\cdot}3{\cdot}2{\cdot}1 = 720$ ways.

3.2.2 Combinations

As noted earlier, the order in which objects are arranged matters in *permutations*. However, order does not matter in *combinations*.

The number of combinations of n different objects taken r at a time is a selection of r out of the n objects with no attention given to the order of arrangement. The number of combinations of n objects taken r at a time is denoted by nC_r $C(n, r)$, $C_{n,r}$ or $\left(\frac{n}{r}\right)$ is $r!$ times smaller than the number of permutations.

Therefore,

$$nC_r = \binom{n}{r} = \frac{nP_r}{r!} = \frac{n!}{(n-r)!r!} \tag{3.46}$$

It should be noted from symmetry that the identity is valid

$$nC_r = nC_{n-r} \tag{3.47}$$

Also $\binom{n}{r} = \frac{nP_r}{r!}$

The arrangement for dividing the number of permutations by $r!$ to get the number of combinations is that each combination give rise to $r!$ possibilities when arranged in all possible ways, thereby giving all the permutations.

Tables of binomial coefficients $\left\{\frac{n}{r}\right\}$ are given in Appendix-*B*.

Note that in combinations, n is always greater than or equal to r. If n is smaller than r, then we cannot select r distinct elements from n.

Example E3.23

(a) Evaluate the following:

$$\binom{8}{4}, \binom{6}{6}, \binom{9}{3}, \binom{4}{0}, \binom{5}{3}, \binom{15}{2} \text{ and } \binom{20}{8}$$

(b) Four cards are chosen in succession from a deck of 52 cards. Find the number of ways this can be done *(i)* with replacement, *(ii)* without replacement.

(c) find the number m of groups of 3 that can be formed from 8 students in a class.

SOLUTION:

(a)

$$\binom{8}{4} = \frac{8\cdot7\cdot6\cdot5}{4!} = \frac{8\cdot7\cdot6\cdot5}{4\cdot3\cdot2\cdot1} = 70$$

$$\binom{6}{6} = \frac{6!}{6!(6-6)!} = \frac{1}{0!} = \frac{1}{1} = 1$$

$$\binom{9}{3} = \frac{9!}{(9-3)!3!} = \frac{9!}{6!3!} = \frac{9\cdot8\cdot7\cdot6\cdot5\cdot4\cdot3\cdot2\cdot1}{6\cdot5\cdot4\cdot3\cdot2\cdot1\cdot3\cdot2\cdot1} = 84$$

$$\binom{4}{0} = \frac{4!}{0!(4-0)!} = \frac{4!}{1\cdot4!} = 1$$

$$\binom{5}{3} = \frac{5!}{(5-3)!3!} = \frac{5!}{2!3!} = \frac{5.4.3.2.1}{2.1.3.2.1} = 10$$

$$\binom{15}{2} = \frac{15!}{(15-2)!2!} = \frac{15\cdot14\cdot13\cdots\cdot2\cdot1}{13\cdot12\cdots\cdot2\cdot1\cdot2\cdot1} = 105$$

$$\binom{20}{8} = \frac{20!}{(20-8)!8!} = 125,970$$

(b) *(i)* Since each card is replaced before the next card is chosen, each card can be chosen by 52 ways. Hence, there are

$(52)(52)(52)(52) = 52^4 = 7,311,616$

different ordered samples of size $n = 4$ with replacement.

(ii) Since there is no replacement, the first card can be chosen in 52 ways, the second card in 51 ways, 3rd card in 50 ways and the last card in 49 ways. Thus there are

$P(52, 4) = 52(51)(50)(49) = 6,497,400$

different ordered samples of size $r = 4$ without replacement.

(c)

$$m = C(8,3) = \binom{8}{3} = \frac{8\cdot7\cdot6}{3\cdot2\cdot1} = 56$$

Since each group is essentially a combination of 8 students taken 3 at a time.

Example E3.24

The president of a university has 12 departments under his administration. The president plans to visit 3 of these departments during the next week. If he randomly selects 3 departments from these 12, how many total selections are possible? Verify the answer using the table in Appendix-*B*.

SOLUTION:

The total possible selections for selecting 3 departments from 12 are

$$\binom{12}{3} = \frac{12!}{(12-3)!3!} = 220$$

From the table in Appendix-*B*, we read for $n = 12$ and $r = 3$, the answer is 220.

Example E3.25

Out of 5 men and 7 women, a committee consisting of 2 men and 3 women is to be formed. In how many ways can this be done if

(*a*) any men and any women can be included

(*b*) one particular women must be included on the committee

(*c*) two particular men cannot be included on the committee.

SOLUTION:

(*a*) 2 men out of 5 can be selected in $5C_2$ ways.

3 women out of 7 can be selected in $7C_3$ ways.

Hence, the number of possible selections = $(5C_2)\,(7C_3) = (10)\,(35) = 350$

(*b*) 2 men out of 5 can be selected $5C_2$ ways.

2 additional women out of 6 can be selected in $6\ C_2$ ways.

Hence, the number of possible selections = $(5C_2)\,(6C_3) = (10)\,(15) = 150$

(*c*) 2 men out of 3 can be selected $3C_2$ ways.

3 women out of 7 can be selected in $7C_2$ ways.

Hence, the number of possible selections = $(3C_2)\,(7C_3) = (3)\,(35) = 105$

3.3 DISCRETE DISTRIBUTIONS

Discrete random variables are used to describe random phenomena in which only integer values can occur.

3.3.1 Hypergeometric Distribution

In this section we consider dependent Bernoulli random variables. A common source of dependent Bernoulli random variables is sampling without replacement from a finite population. Suppose that a finite population consists of a known number of successes and failures. If we sample a fixed number of units from that population, the number of successes in our sample will have a distribution that is a member of the family of hypergeometric distributions.

Definition of the Hypergeometric Distribution

A set of N objects contains K objects classified as successes and $N - K$ objects classified as failures. A sample of size n objects is selected randomly (without replacement) from the N objects, where $K \leq N$ and $n \leq N$.

Let the random variable X denotes the number of successes in the sample. Then X is a *hypergeometric random variable* and

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}} \qquad x = \max\{0, n+K-N\} \text{ to } \min\{K, n\} \tag{3.48}$$

The expression $\{K, n\}$ is used in the definition of the range of X because the maximum number of successes that can occur in the sample is the smaller of the sample size, n, and the number of successes available, K.

Also, if $n + K > N$, at least $n + K - N$ successes must occur in the sample.

It should be noted here that in Eq. (3.48)

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} \tag{3.49}$$

is the number of a parts taken b at a time. The hypergeometric distribution is the appropriate possibility model for sampling without replacement. Other distributions are often employed when the ratio of n/N becomes small, say 0.10 or less.

The mean and variance of a hypergoemetric random variable can be determined from the trials that comprise the experiment. However, the trials are not independent, and so the calculations are more difficult than for a binomial distribution. The results are stated as follows:

Mean and Variance of a Hypergeometric Distribution

If X is a hypergoemetric random variable with parameters N, K, and n, then the mean

$$\mu = E(X) = np \tag{3.49a}$$

and the variance

$$\sigma^2 = V(X) = np(1-p)\left(\frac{N-n}{N-1}\right) \tag{3.49b}$$

where

$$p = \frac{K}{N}. \tag{3.49c}$$

Here, p is interpreted as the proportion of successes in the set of N objects. For a hypergeometric random variable, $E(X)$ is similar to the mean of a binomial random variable. Also, $V(X)$ differs from the result for a binomial random variable only by the term shown below:

The term in the variance of a hypergeometric random variable $\frac{N-n}{N-1}$ is called the *finite population correlation factor.*

Example E3.26

Given that X has a hypergeometric distribution with $N = 100$, $n = 4$ and $K = 20$. Determine the following:

(*a*) $P(X = 1)$

(*b*) $P(X = 6)$

(*c*) $P(X = 4)$

(*d*) the mean and variance of X.

SOLUTION:

Here, $K = 20$, $X = 1$, $N = 100$ and $n = 4$.

$$(a)\quad P(X=1) = \frac{\binom{K}{n}\binom{N-K}{n-X}}{\binom{N}{n}} = \frac{\binom{20}{4}\binom{80}{3}}{\binom{100}{4}} = \frac{20(82160)}{3921225} = 0.4191$$

(*b*) $P(X = 6) = 0$, since the sample size is only 4.

$$(c)\quad P(X=4) = \frac{\binom{K}{n}\binom{N-K}{n-X}}{\binom{N}{n}} = \frac{\binom{20}{4}\binom{80}{0}}{\binom{100}{4}} = \frac{4845(1)}{3921225} = 0.001236$$

$$(d)\quad \text{Mean} = E(X) = np = n\left(\frac{K}{N}\right) = 4\left(\frac{20}{100}\right) = 0.8$$

$$\text{Variance} = V(X) = np(1-p)\left(\frac{N-n}{N-1}\right) = 40(0.2)(0.8)\left(\frac{96}{99}\right) = 0.6206$$

Example E3.27

Suppose a box contains five red balls and ten blue balls. If seven balls are selected at random without replacement, find the probability that at least 3 red balls will be obtained.

SOLUTION:

Let X denotes the number of red balls that are obtained. Then, X has a hypergeometric distribution, $D = 5$, $A + B = 15$, $n = 7$, $B = 10$ and $A = 5$. The maximum value of X is $\min\{n, A\} = 5$. Hence,

$$P(X \geq 3) = \sum_{x=3}^{5} \frac{\binom{5}{x}\binom{10}{7-x}}{\binom{15}{7}} = \frac{2745}{6435} = 0.4266$$

Example E3.28

A lot of 75 gaskets contains five in which the variability in thickness around the circumference of the gasket is unacceptable. A sample of 10 gaskets is selected at random, without replacement. What is the probability that

(a) none of the unacceptable gaskets is in the sample

(b) at least one unacceptable gasket is in the sample

(c) exactly one unacceptable gasket is in the sample

(d) the mean number of unacceptable gaskets in the sample.

Solution:

Let X denotes the number of unacceptable gaskets in the sample of 10.

$$f(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

(a) $$P(X=0) = \frac{\binom{5}{0}\binom{70}{10}}{\binom{75}{10}} = \frac{\frac{70!}{10!60!}}{\frac{75!}{10!65!}} = \frac{65\times 64\times 63\times 62\times 61}{75\times 74\times 73\times 72\times 71} = 0.4786$$

(b) $P(X \geq 1) = 1 - P(X=0) = 1 - 0.4786 = 0.5214$

(c) $$P(X=0) = \frac{\binom{5}{1}\binom{70}{9}}{\binom{75}{10}} = \frac{\frac{5!70!}{9!6!}}{\frac{75!}{10!65!}} = \frac{5\times 65\times 64\times 63\times 62\times 10}{75\times 74\times 73\times 72\times 71} = 0.3923$$

(d) $$E(X) = 10\left(\frac{5}{75}\right) = \frac{2}{3}$$

Example E3.29

Of 50 manufactured steel rods in a production process by a company, 12 have defects. If 10 steel rods are selected at random for inspection,

(a) find the probability that exactly 3 of the 10 have defects

(b) find the mean and variance of X.

Solution:

(a) Let X represents the number of sampled steel rods that have defects. Then, $K = 12$, $N = 50$ and $n = 10$.

$$P(X=3)=\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}=\frac{\binom{12}{3}\binom{38}{7}}{\binom{50}{10}}=\frac{(220)(12,620,256)}{10,272,278,170}=0.2703$$

(b) Mean = $\mu_X = np = \dfrac{(10)(12)}{50} = 2.4$

where $p = \dfrac{12}{50}$

Variance = $\sigma_X^2 = np(1-p)\left(\dfrac{N-n}{N-1}\right) = 10\left(\dfrac{12}{50}\right)\left(1-\dfrac{12}{50}\right)\left(\dfrac{50-10}{50-1}\right) = 1.4890$

Example E3.30

A large bin contains 80 balls of which 32 are red balls and 48 are blue balls. Suppose 15 balls are picked at random. Find the probability of getting 4 red balls, the mean number of red balls, and the standard deviation of the number of red balls if the sample is picked

(a) with replacement

(b) without replacement.

SOLUTION:

We will identify success with "picking a red ball" and let X = number of red balls in the sample.

(a) If the sampling is with replacement, we have a binomial distribution with 15 trials and probability of success $p = \dfrac{32}{80} = 0.4$. Hence,

$P(X=4)$ = Binomial(4; 15, 0.4) = 0.127 (from binomial distribution in Appendix-*C*)

$\mu = np = 15(0.4)$

$\sigma = \sqrt{np(1-p)} = \sqrt{15(0.4)(0.6)} = 1.9$

(b) If the sampling is without replacement, we have a hypergeometric distribution and

$$P(X=4)=\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}=\frac{\binom{32}{4}\binom{48}{11}}{\binom{80}{15}}=0.122$$

Note that this answer is not much different from the one obtained in part *(a)* using the binomial formula. The reason is that N (= 80), compared to n (= 15), is fairly large.

Mean, $\mu = np = 15(0.4) = 6$

Standard deviation

$$\sigma = \sqrt{np(1-p)\left(\frac{N-n}{n-1}\right)} = \sqrt{15\left(\frac{32}{80}\right)\left(1-\frac{32}{80}\right)\left(\frac{80-15}{80-1}\right)} = 1.72$$

3.3.2 The Binomial Probability Distribution

The binomial probability distribution is one of the most widely used discrete probability distribution. It is used to find the probability that an outcome will occur x times in n performances of an experiment. The binomial distribution is applied to experiments that satisfy the four conditions of a *binomial experiment*. Each repetition of a binomial experiment is called a *trial* or a *Bernoulli trial* (after Jacob Bernoulli).

A trial with only two possible outcomes is used so frequently as a building block of a random experiment that is called a *Bernoulli trial*. It is usually assumed that the trials that constitute the random experiment are *independent*. This implies that the outcome from one trial has no effect on the outcome to be obtained from any other trial. In addition, it is often reasonable to assume that the *probability of a success in each trial is constant*.

3.3.3 The Binomial Experiment

An experiment that satisfies the following four conditions is called a *binomial experiment*.

1. There are n identical trials. In other words, the given experiment is repeated n times. All these repetitions are performed under identical conditions.
2. Each trial has two and only two outcomes. These outcomes are usually called a *success* and a *failure*.
3. The probability of success is denoted by p and that of failure by q, and $p + q = 1$. The probabilities p and q remain constant for each trial.
4. The trials are independent. In other words, the outcome of one trial does not affect the outcome of another trial.

One of the two outcomes of a trial is called a *success* and the other a *failure*.

The random variable X that equals the number of trials that result in a success has a *binomial random variable* with parameters $0 < p < 1$ and $n = 1, 2, \ldots$. The probability mass function (pmf) of X is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, 2, \ldots, n \tag{3.50}$$

3.3.4 The Binomial Formula

3.3.4.1 Binomial Theorem

The n^{th} power of $(p + q)$ can be expressed in terms of binomial coefficients as

$$(p+q)^n = p^n + np^{n-1}q + \frac{n(n-1)}{2!}p^{n-2}q^2 + \cdots + \frac{n!}{r!(n-r)!}p^{n-r}q^r + \cdots + q^n$$

$$= \sum_{r=0}^{n}\binom{n}{r}p^{n-r}q^r = \sum_{r=0}^{n} {}_nC_r\, p^{n-r}q^r \tag{3.50a}$$

Since $(p + q) = (q + p)$, we can also write

$$(p+q)^n = \sum_{r=0}^{n} {}_nC_r p^r q^{n-r} \tag{3.51}$$

Consider n trials in each of which the probability of success is p. Then the probability of failure is $1 - p = q$. To find the probability of r successes, we observe that:

the probability of 1 success in 1 try is p,

the probability of 2 successes in 2 tries is $p \times p$ or p^2,

the probability of 3 successes in 3 tries is $p \times p \times p$ or p^3,

$\vdots$

the probability of r successes in r tries is p^r and

the probability of subsequent $(n - r)$ failures in $(n - r)$ tries is $(1 - p)^{n-r} = q^{n-r}$.

The probability of r successes followed by $(n - r)$ failures is $p^r(1 - p)^{n\ r}$. Here, we have considered only one particular group or combination of r events; i.e., we have started with r successes and finished with $(n - r)$ failures; every other possible ordering of r successes and $(n - r)$ failures will also have the same probability.

The number of possible orderings or the number of selections for r successes and $(n - r)$ failures in n trials is $n!/[r!(n - r)!]$.

Therefore, the probability P_r of an event succeeding r times is

$$P_r = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \tag{3.52}$$

or

$$P_r = {}_nC_r p^r q^{n-r} \tag{3.53}$$

This term is similar to the r^{th} term of the binomial expansion $(q + p)^n$ [see Eq. (3.51)], which can be written

$$(q+p)^n = \sum_{r=0}^{n} {}_nC_r p^r q^{n-r} \tag{3.54}$$

The successive terms of the expansion give the probability P_r of an event succeeding r times in n trials for values of r varying in steps of one of 0 to n.

3.3.4.2 Cumulative Terms for Binomial Distribution

The probability of an event succeeding at least r times in n trials.

The probability P_r of an event succeeding exactly r times in n trials is given by Eq. (3.53)

$$P_r = {}_nC_r p^r q^{n-r}$$

The probability of an event succeeding at least r' times in n trials is given by

$$\sum_{r=r'}^{r=n} P_r = \sum_{r=r'}^{r=n} {}_nC_r p^r q^{n-r} \tag{3.55}$$

The values of the summation of Eq. (3.55) are given in Appendix-*D* for p ranging from 0.05 to 0.50; with n between 2 and 20, and r' between 1 and 20.

For $p > 0.5$, we can utilise the fact that the probability is

$$P_r = 1 - \sum_{r=n-r'+1}^{r=n} {}_nC_r q^r p^{n-r} \tag{3.56}$$

3.3.4.3 Mean and Standard Deviation of Binomial Distribution

If p is the proportion of successes in the population, then the mean number of successes in n trials is

$$\mu = np \tag{3.57}$$

This is clear, as the mean number of successes in n trials is equal to the probability of success in one trial times the number of trials.

The standard deviation for a binomial frequency distribution is

$$\sigma = \sqrt{npq} \text{ or } \sqrt{np(1-p)} \tag{3.58}$$

q is not independent but is equal to $(1 - p)$. The binomial distribution can be expressed in terms of two parameters, n and p.

Equations (3.57) and (3.58) will now be derived using the definition of expectation for the mean and variance. We have

$$E(r) = \mu = \sum_{r=0}^{n} rP_r = \sum_{r=0}^{n} r\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}$$

$$= np\sum_{r=1}^{n}\frac{(n-1)!}{(r-1)!(n-r)!}p^{r-1}(1-p)^{n-r} = np[p+q]^{n-1}$$

or

$$\mu = np \tag{3.59}$$

For the variance σ^2

$$E(r-\mu)^2 = \sigma^2 = \sum_{r=0}^{n}(r-\mu)^2 P_r = \sum_{r=0}^{n} r^2 P_r - 2\mu\sum_{r=0}^{n} rP_r + \mu^2\sum_{r=0}^{n} P_r$$

$$= \sum_{r=0}^{n} r^2 P_r - 2\mu(\mu) + \mu^2$$

Since

$$\sum_{r=0}^{n} P_r = 1$$

and

$$\sum_{r=0}^{n} rP_r = \mu = np$$

Thus

$$\sigma^2 \sum_{r=0}^{n} r^2 P_r - (np)^2 = \sum_{r=0}^{n}[r(r-1)+r]P_r - (np)^2$$

$$= \sum_{r=2}^{n} r(r-1)\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r} + \sum_{r=0}^{n} rP_r - (np)^2$$

$$= n(n-1)p^2 \sum_{r=2}^{n} \frac{(n-2)!}{(r-2)!(n-r)!} p^{r-2}(1-p)^{n-r} + np - (np)^2$$

$$= n(n-1)p^2 (p+q)^{n-2} + np - (np)^2$$

or $$\sigma^2 = n(n-1)p^2 + np - (np)^2 = np(1-p) = npq \tag{3.60}$$

Hence $$\sigma = \sqrt{npq} \tag{3.61}$$

Summarizing, if X is a binomial random variable with parameters p and n, then the mean,

$$\mu = E(X) = np \tag{3.61a}$$

and the variance,

$$\sigma^2 = V(X) = np(1-p) \tag{3.61b}$$

A binomial distribution with $n = 20$ and $p = 0.10$ is shown in Fig. 3.4(a).

Cumulative binomial distribution are given in Appendix-D.

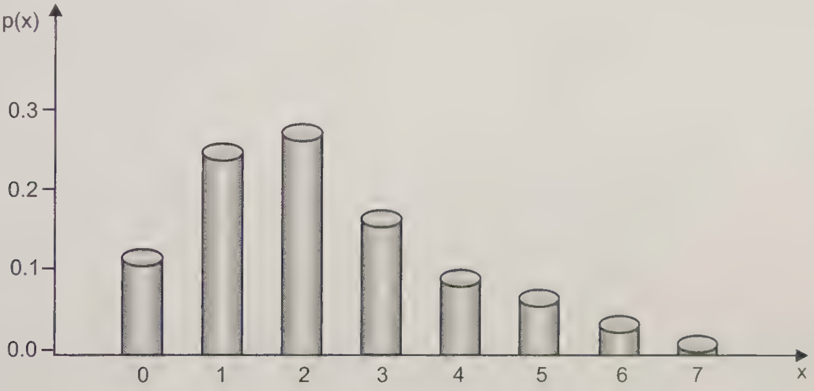


Fig. 3.4(a): Binomial pmf for n = 20, p = 0.10

Example E3.31

Let X denotes the number of mechanical components that are defective in a testing process and assume that X is a binomial random variable with $p = 0.001$. If 1000 of these components are tested, find the following:

(a) $P(X = 1)$

(b) $P(X \geq 1)$

(c) $P(X \leq 2)$

(d) mean and variance of X

SOLUTION:

We have from Eq. (3.50)

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

(a) $P(X = 1) = \binom{1000}{1} (0.001)^1 (0.999)^{999} = 0.368$

(b) $$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{1000}{1}(0.001)^1(0.999)^{999} = 0.632$$

(c) $$P(X \leq 2) = \binom{1000}{0}(0.001)^0(0.999)^{1000} + \binom{1000}{1}(0.001)^1(0.999)^{999}$$

$$+ \binom{1000}{2}(0.001)^2(0.999)^{998} = 0.920$$

(d) $E(X) = 1000(0.001) = 1$

$V(X) = 1000(0.001)\,(0.999) = 0.999$

Example E3.32

A professional basket player makes 80% of the free throws he tries. Assuming this percentage will hold true for future attempts, find the probability that in the next eight tries the number of free throws he will make is

(a) exactly 8

(b) exactly 5.

SOLUTION:

Here $n = 8$, $p = 0.80$, $q = 1 - p = 1 - 0.80 = 0.20$

(a) $$P(\text{exactly } 8) = P(8) = \binom{n}{x} p^x(1-p)^{n-x} = \binom{8}{8}(0.80)^8(0.20)^0 = 1(0.167772)(1) = 0.1678$$

(b) $$P(\text{exactly } 5) = P(5) = \binom{n}{x} p^x(1-p)^{n-x} = \binom{8}{5}(0.80)^5(0.20)^3 = (56)(0.32768)(0.008) = 0.1468$$

Example E3.33

A soft drink company conducted a taste survey before marketing a new soft drink. The results of the survey showed that 80% of the people like this soft drink. On a certain day, 8 customers bought it.

(a) let x denotes the number of customers in this sample of 8 who will like this soft drink. Using the binomial probabilities table, find the probability distribution of x and draw a graph of the probability distribution. Find the mean and standard deviation.

(b) Using the binomial distribution of part *(a)*, find the probability that exactly three of the eight customers will like the soft drink.

SOLUTION:

(a) Here $n = 8$, $p = 0.80$. From Appendix-*C*, for $n = 8$, and $p = 0.88$ we have x, $P(x)$ tabulated as in Table E3.33.

Table E3.33

x	P(x)
0	0.0000
1	0.0001
2	0.0011
3	0.0092
4	0.0459
5	0.1468
6	0.2936
7	0.3355
8	0.1678

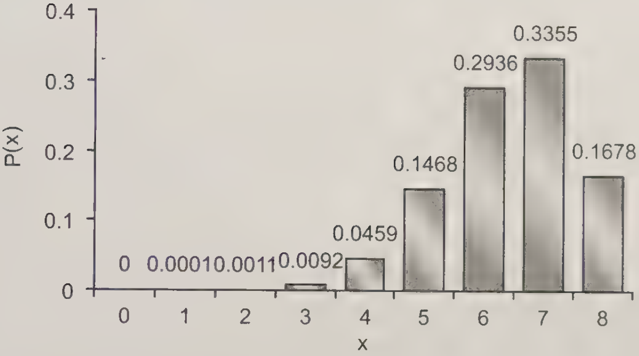


Fig. E3.33

The mean and standard deviation of x are (from Eqs. (3.61a) and (3.61b)

Mean

$$\mu = np = 8(0.80) = 6.4$$

Standard deviation

$$\sigma = \sqrt{npq} = \sqrt{8(0.80)(0.20)} = 1.1317$$

(*b*) P(exactly 3 customers like the soft drink) = $P(3) = 0.0092$

Example E3.34

A certain mechanical system contains 10 components. Assuming that the probability of each individual component will fail is 0.2 and the components fail independently of each other. Given that at least one of the components has failed, what is the probability that at least 2 of the components have failed?

SOLUTION:

The number X of components that fail will have a binomial distribution with parameters $n = 10$ and $p = 0.2$. Hence,

$$P(X \geq 2 \mid X \geq 1|) = \frac{P(X \geq 2)}{P(X \geq 1)} = \frac{1 - P(X = 0) - P(X = 1)}{1 - P(X = 0)}$$

From Appendix-*C*, for $n = 10, p = 0.2$, we have

$$P(X \geq 2 \mid X \geq 1|) = \frac{1-0.1074-0.2684}{1-0.1074} = \frac{0.6242}{0.8926} = 0.6993$$

Example E3.35

(a) If a fair coin is (probability of heads equals 1/2) is tossed independently 10 times. Use the table in Appendix-*C* of the binomial distribution to find the probability that strictly more heads are obtained than tails.

(b) Suppose that the probability that a certain experiment will be successful is 0.3, and let X denotes the number of successes that are obtained in 15 independent performances of the experiment. Use the table in Appendix-*C* for the binomial distribution to determine the value of $P(6 \leq X \leq 9)$.

SOLUTION:

(a) Let X be the number of heads obtained. More heads than tails are obtained if $X \in \{6, 7, 8, 9, 10\}$. The probability of this event is the sum of the numbers in the binomial table in Appendix-*C*. Corresponding to

$$P = 0.5$$

and $n = 10$

for $x = 6, 7, 8, 9, 10 = 0.2051 + 0.1172 + 0.0439 + 0.0098 + 0.0010 = 0.37695.$

By the symmetry of this binomial distribution, we can also compute the sum as

$$(1 - P(X = 5))/2 = (1 - 0.2461)/2 = 0.37695.$$

(b) From the table in Appendix-*C* for the binomial distribution with parameters $n = 15$ and $p = 0.3$, that

$$\begin{aligned} P(6 \leq X \leq 9) &= P(X = 6) + P(X = 7) + P(X = 8) + P(X = 9) \\ &= 0.1472 + 0.0811 + 0.0348 + 0.0116 = 0.2747 \end{aligned}$$

Example E3.36

According to a particular survey, the probability is 0.40 that a traffic fatality involves an intoxicated or alcohol-impaired driver or non-occupant. In 8 traffic fatalities, find the probability that the number, A, which involve an intoxicated or alcohol-impaired driver or non-occupant is

(a) exactly 3; at least 3; at most 3

(b) between 2 and 4, inclusive

(c) find and interpret the mean of the random variable A

(d) obtain the standard deviation of A.

SOLUTION:

Here, $n = 8$ and $p = 0.40$. Thus $q = 1 - p = 1 - 0.40 = 0.60$

From Eq. (3.50), we have

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Refer to the table in Appendix-*C* for $n = 8$ and $p = 0.40$

(*a*) $P(\text{exactly } 3) = \binom{8}{3}(0.4)^3(0.6)^5 = 0.0241$

$$P(3 \text{ or more}) = 1 - P(0 \text{ or } 1 \text{ or } 2) = 1 - \binom{8}{0}(0.4)^0(0.6)^8 - \binom{8}{1}(0.4)^1(0.6)^7 - \binom{8}{2}(0.4)^2(0.6)^6$$

$$= 1 - 0.0168 - 0.0896 - 0.2090 = 0.6846$$

$P(\text{at most } 3) = P(0) + P(1) + P(2) + P(3)$

$$= \binom{8}{0}(0.4)^0(0.6)^8 - \binom{8}{1}(0.4)^1(0.6)^7 - \binom{8}{2}(0.4)^2(0.6)^6 + \binom{8}{3}(0.4)^3(0.6)^5$$

$$= 0.0168 + 0.0896 + 0.2090 + 0.2787 = 0.5941$$

(*b*) $P(2 \le A \le 4) = P(2) + P(3) + P(4) = \binom{8}{2}(0.4)^2(0.6)^6 - \binom{8}{3}(0.4)^3(0.6)^5 - \binom{8}{4}(0.4)^4(0.6)^4$

$$= 0.2092 + 0.2787 + 0.2322 = 0.7201 \approx 0.72$$

(*c*) The mean of A is $\mu = np = 8(0.4) = 3.2$. On average, of 8 traffic fatalities, 3.2 will involve an intoxicated or alcohol-impaired driver or non-occupant.

(*d*) Standard deviation $= \sqrt{\sigma^2} = \sqrt{np(1-p)} = \sqrt{8(0.4)(0.6)} = \sqrt{19.2} = 4.3418$

Example E3.37

According to a particular survey, 14.9% of those who have received a doctor's degree in engineering are blacks. Suppose that 6 people who have received their doctor's degree in engineering are randomly selected. Find the probability that

(*a*) exactly 2 are black

(*b*) exactly 4 are black

(*c*) at least 2 are black

(*d*) find the probability distribution of the number of blacks in a sample of 6 persons who have received their doctor's degree in engineering

(*e*) why is the probability distribution obtained in part (*d*) only approximately correct? What is the exact distribution called?

SOLUTION:

Here, $n = 6$ and $p = 0.149$ and $q = 1 - p = 1 - 0.149 = 0.851$

From Eq. (3.50), we have

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

(*a*) $P(2) = \binom{6}{2}(0.149)^2(0.851)^4 = 0.1747$

(*b*) $P(4) = \binom{6}{4}(0.149)^4(0.851)^2 = 0.0054$

(*c*) $P(\text{at least } 2) = 1 - P(0 \text{ or } 1) = 1 - \binom{6}{0}(0.149)^0(0.851)^6 - \binom{6}{1}(0.149)^1(0.851)^5$

$$= 1 - 0.3798 - 0.3990 = 0.2212$$

(*d*) $P(X = x) = \binom{6}{x}(0.149)^x(0.851)^{6-x}$ for $x = 0, 1, 2, \ldots, 6.$

Applying this formula for each value of *x* gives the results in the Table E3.37.

Table E3.37

x	P(X = x)
0	0.3798
1	0.3990
2	0.1747
3	0.0408
4	0.0054
5	0.0004
6	0.0000

(*e*) The sampling was actually done without replacement, so the trials are not independent and the success probability changes very slightly from trial to trial. The exact probability distribution is called a hypogeometric distribution.

Example E3.38

Suppose the probability is 0.67 that the favorite in a horse race will finish in the money (first, second, or third place). In the next 5 races, what is the probability that the favorite finishes in the money

(*a*) exactly 2 times

(*b*) exactly 4 times

(*c*) at least 4 times

(*d*) between 2 and 4 times, inclusive

(*e*) find the probability distribution of the random variable, *X*, the number of times the favorite finishes in the money in the next 5 races

(*f*) identify the probability distribution of *X* as right skewed, symmetric, or left skewed without checking its probability distribution or its histogram

(*g*) Draw a probability histogram for *X*

(*h*) find the mean and standard deviation of the random variable *X*.

SOLUTION:

Refer to the table in Appendix-*C*.

Here $n = 5, p = 0.67$ and $q = 1 - p = 1 - 0.67 = 0.33$

From Eq. (3.50) we have

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$P(0) = \binom{5}{0} (0.67)^0 (0.33)^5 = 0.004$$

$$P(1) = \binom{5}{1} (0.67)^1 (0.33)^4 = 0.040$$

$$P(2) = \binom{5}{2} (0.67)^2 (0.33)^3 = 0.161$$

$$P(3) = \binom{5}{3} (0.67)^3 (0.33)^2 = 0.328$$

$$P(4) = \binom{5}{4} (0.67)^4 (0.33)^1 = 0.332$$

$$P(5) = \binom{5}{5} (0.67)^5 (0.33)^0 = 0.135$$

(a) $P(2) = 0.161$

(b) $P(4) = 0.332$

(c) $P(X \geq 4) = P(4) + P(5) = 0.332 + 0.135 = 0.467$

(d) $P(2 \leq X \leq 4) = P(2) + P(3) + P(4) = 0.161 + 0.328 + 0.332 = 0.821$

(e)

x	P(X = x)
0	0.004
1	0.040
2	0.161
3	0.328
4	0.332
5	0.135

(f) Left-skewed

(g) See Fig. E3.38.

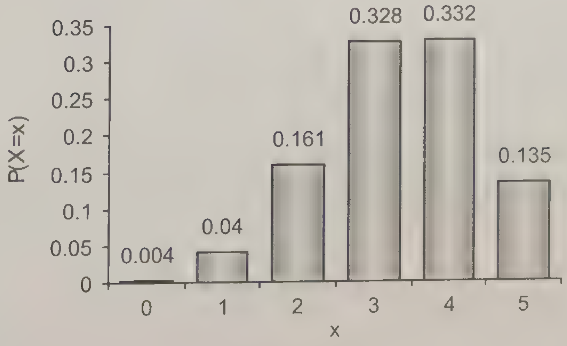


Fig. E3.38

x	P(X = x)	xP(X = x)	x^2	x^2P(X = x)
0	0.004	0.000	0	0.000
1	0.040	0.040	1	0.040
2	0.161	0.322	4	0.644
3	0.328	0.984	9	2.952
4	0.332	1.328	16	5.312
5	0.135	0.675	25	3.375
Σ		3.349		12.323

(h) Mean = $\mu = 3.349$ (from the above table)

Variance,

$$\sigma^2 = 12.323 - 3.349^2 = 1.107$$

Hence, standard deviation = $\sqrt{1.107} = 1.052$

Alternate calculation:

$$\mu = np = 5(0.67) = 3.35$$

$$\sigma^2 = np(1-p) = 5(0.67)(0.33) = 1.1055$$

or

$$\sigma = \sqrt{1.1055} = 1.051$$

3.3.5 Poisson Distribution

The Poisson distribution, named after the French mathematician Simeon D. Poisson is another important probability distribution of a discrete random variable that has many applications. The Poisson distribution is applied to experiments with random and independent occurrences.

Given an interval of real numbers, assume events occur at random throughout the interval. If the interval can be partitioned into subintervals of small enough length such that

1. the probability of more than one event in a subinterval is zero
2. the probability of one event in a subinterval is the same for all subintervals and proportional to the length of the subinterval, and
3. the event in each subinterval is independent of other subintervals, the random experiment is called a *Poisson process*.

Independence of occurrences means that one occurrence (or non-occurrence) of an event does not influence the successive occurrence or nonoccurrences of that event. The occurrences are always considered with respect to an interval. The interval may be a time interval, a space interval, or a volume interval. The actual number of occurrences within an interval is random and independent. If the average number of occurrences for a given interval is known, then by using the Poisson probability distribution we can determine the probability of a certain number of occurrences, x, in that interval. Note that the number of actual occurrences in an interval is denoted by x.

Conditions to apply Poisson probability distribution

The following three conditions must be satisfied to apply the Poisson probability distribution:

1. x is a discrete random variable
2. The occurrences are random
3. The occurrences are independent

The following are a few examples of discrete random variables for which the occurrences are random and independent.

Some of the phenomena that follow the Poisson distribution are:

1. counts of flaws in castings
2. the number of vehicles on a highway
3. the number of customers visiting a bank
4. the number of accidents that occur on a given highway during a period of time
5. the number of telephone calls
6. counts of power outages
7. counts of atomic particles emitted from a specimen.

The random variable X that equals the number of events in the interval is a *Poisson random variable* with parameter $0 < \lambda$, and the probability mass function (pmf) of X is

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}; \qquad x = 0, 1, 2, \ldots \tag{3.62}$$

The Poisson distribution is made up of a series of terms:

$$e^{-1}, e^{-1}\lambda, \frac{e^{-\lambda}\lambda^{2}}{2!}, \frac{e^{-\lambda}\lambda^{3}}{3!}, \ldots.$$

representing respectively, the probabilities of occurrence of 0, 1, 2, 3, 4, etc. events, where e is the base of the natural logarithm and λ is the mean frequency of occurrence. The sum of the probabilities is one (1) because

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^{k}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{k}}{k!} \tag{3.62a}$$

and the summation on the right-hand side of the equation (3.62*a*) is recognized to be Taylor's expansion of e^{x} evaluated at λ. Hence, the summation equals e^{λ} and the right-hand side equals $e^{-\lambda}e^{\lambda} = 1$. Poisson distribution for $\lambda = 3$ is shown in Fig. 3.4(*b*).

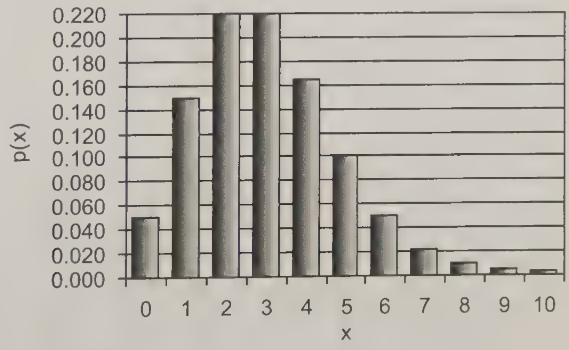


Fig. 3.4(b)

3.3.5.1 Derivation from Binomial Distribution

The Poisson distribution can also be deduced from the binomial distribution, provided that n is large ($\to\infty$), p is very small ($\to 0$), and np is finite. It was shown earlier that in n trials the probability of an event succeeding r times is

$$P_r = \frac{n!}{r!(n-r)!} p^r q^{n-r} \tag{3.63}$$

When n is large compared with r,

$$\frac{n!}{(n-r)!} = n(n-1)(n-2)\ldots(n-r+1) \simeq n^r$$

Therefore, the probability of r successes becomes

$$P_r = \frac{n^r}{r!} p^r q^{n-r} \tag{3.64}$$

Now if p is very small and r is not large,

$$q^r = (1-p)r \simeq 1$$

and $$q^{n-r} \simeq q^n = (1-p)^n$$

Hence $$P_r = \frac{(np)^r}{r!}(1-p)^n = \frac{(np)^r}{r!}\left[1 - np + \frac{n(n-1)(-p)^2}{2!} + \frac{n(n-1)(n-2)(-p)^3}{3!} + \cdots\right]$$

$$\simeq \frac{(np)^r}{r!}\left[1 - np + \frac{(np)^2}{2!} - \frac{(np)^3}{3!} + \cdots\right]$$

Thus $$P_r = \frac{(np)^r}{r!} e^{-np} \tag{3.65}$$

This, then, is the probability of r successes in n trials.

3.3.5.2 Mean and Standard Deviation

The mean number of occurrences of an event per unit of time (or space) is

$$\mu = np \tag{3.66}$$

and the standard deviation of the numbers of events is

$$\sigma = \sqrt{np} \tag{3.67}$$

Thus the mean and variance are equal to one another:

$$\mu = \sigma^2 = np \tag{3.68}$$

Equations (3.66) and (3.68) will now be derived.

$$\text{mean} = E(r) = \sum_{r=0}^{\infty} rP_r = \sum_{r=0}^{\infty} \frac{re^{-\lambda}\lambda^r}{r!} = 0 + \lambda e^{-\lambda} + 2\frac{\lambda^2 e^{-\lambda}}{2!} + 3\frac{\lambda^3 e^{-\lambda}}{3!} + \cdots$$

$$= \lambda e^{-\lambda}\left[1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right] = \lambda e^{-\lambda} e^{\lambda} = \lambda \tag{3.69}$$

The variance σ^2 is given by

$$\sigma^2 = E(r-\lambda)^2 = \sum_{r=0}^{\infty}(r-\lambda)^2 P_r$$

$$\sigma^2 = \sum_{r=0}^{\infty} r^2 P_r - 2\lambda\sum_{r=0}^{\infty} rP_r + \lambda^2\sum_{r=0}^{\infty} P_r$$

Now $$-2\lambda\sum_{r=0}^{\infty} rP_r = -2\lambda(\lambda) = -2\lambda^2$$

$$\lambda^2\sum_{r=0}^{\infty} P_r = \lambda^2(1) = \lambda^2$$

and $$\sum_{r=0}^{\infty} r^2 P_r = \sum_{r=0}^{\infty}[r(r-1)+r]P_r = \sum_{r=0}^{\infty} r(r-1)P_r + \lambda = \left[0+0+2\frac{\lambda^2 e^{-\lambda}}{2!}+6\frac{\lambda^3 e^{-\lambda}}{3!}+\cdots\right]+\lambda$$

$$= \lambda^2 e^{-\lambda}\left[1+\lambda+\frac{\lambda^2}{2!}+\cdots\right]+\lambda = \lambda^2 e^{-\lambda}(e^{\lambda})+\lambda = \lambda^2+\lambda$$

Thus, collecting the terms, we have

$$\sigma^2 = \lambda^2+\lambda-2\lambda^2+\lambda^2$$

or $$\sigma^2 = \lambda \tag{3.70}$$

Hence, if X is a Poisson random variable with parameter λ, then

the mean $\mu = E(X) = \lambda$, and

the variance $\sigma^2 = V(X) = \lambda^2$

Note that the Poisson distribution contains only one parameter, *np*, the mean occurrence of an event, and we do not know the value of *n*. In the binomial distribution we know the number of times an event occurs and the number of times an event does not occur.

Cumulative Poisson probabilities are given in Appendix-*D*. A plot of the Poisson distribution for $\lambda = 3$ is shown in Fig. 3.5. This Poisson distribution has a very long tail to the right (the distribution is skewed).

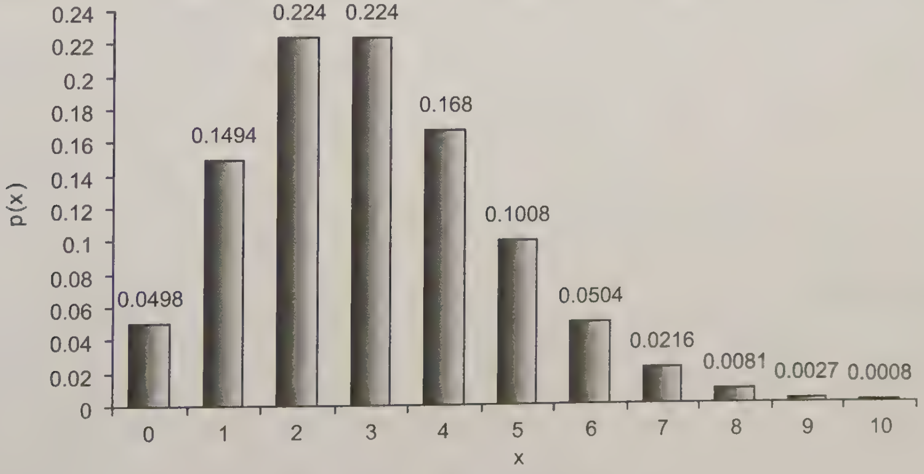


Fig. 3.5: Poisson distribution for $\lambda = 3$

Example E3.39

A survey found that 1.5% of occupied housing units have 7 or more people living within. Use the Poisson distribution to determine the approximate probability that, of 200 randomly selected occupied housing units, there are

(*a*) none with 7 or more persons

(*b*) 3 or more with 7 or more persons

SOLUTION:

We use $\lambda = np = 200(0.15) = 3.0$. We also note that $n \geq 100$ and $np \leq 10$. Hence, apply Eq. (3.62),

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

(*a*) $P(X = 0) = \dfrac{e^{-3.0}(3.0)^0}{0!} = 0.0498$

(*b*) $P(\text{at least } 3) = 1 - P(X = 0 \text{ or } 1 \text{ or } 2)$

$$= 1 - \left[e^{3}\frac{3^0}{0!} + e^{-3}\frac{3^1}{1!} + e^{-3}\frac{3^2}{2!}\right] = 1 - [0.0498 + 0.1494 + 0.2240] = 0.5768$$

Example E3.40

Assume the number of errors along a magnetic recording surface is a Poisson random variable with a mean of one error every 10^5 bits. A sector of data consists of 5000 eight-bit bytes. Find

(*a*) the probability of more than one error in a sector

(*b*) the mean number of sectors until an error is found.

SOLUTION:

Let x denotes the number of errors in a sector. Then, X is a Poisson random variable with $\lambda = 0.4$.

(*a*) $P(X > 1) = 1 - P(X < 1) = 1 - e^{-0.4} - e^{-0.4}(0.4) = 1 - 0.6703 - 0.2681 = 0.06157$

(*b*) Let Y denotes the number of sectors until an error is found. Then, Y is a geometric random variable and

$P = P(X \geq 1) = 1 - P(X = 0) = 1 - e^{-0.4} = 1 - 0.6703 = 0.3297$

$$E(Y) = \frac{1}{p} = \frac{1}{0.3297} = 3.03306$$

Example E3.41

The probability that an individual recovers from an illness in a 2-week time period without medical treatment is 0.1. If 20 independent individuals suffering from this illness are treated with a medicine and 4 recover in a 2-week period. If the medicine has no effect, what is the probability that 4 or more people recover in a 2-week time period?

SOLUTION:

Let x denotes the number of individuals that recover in 2-week time period. Assume the individuals are independent, then, X is binomial random variable with $n = 20$ and $p = 0.1$.

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - [P(X=0) + P(x=1) + P(x=2) + P(X=3)]$$

$$= 1 - \left[\binom{20}{0}(0.1)^0(0.9)^{20} + \binom{20}{1}(0.1)^1(0.9)^{19} + \binom{20}{2}(0.1)^2(0.9)^{18} + \binom{20}{3}(0.1)^3(0.9)^{17}\right]$$

$$= 1 - [0.1216 + 0.2702 + 0.2852 + 0.1901] = 0.1330$$

Example E3.42

A survey found that the traffic flowing through an intersection with an average of 3 cars per 30 seconds. Assume the traffic flow can be modeled as a Poisson distribution.

(a) find the probability of no cars through the intersection within 30 seconds

(b) find the probability of 3 or more cars through the intersection within 30 seconds

(c) find the minimum number of cars through the intersection so that the probability of this number or fewer cars in 30 seconds is at least 90%

(d) if the variance of the number of cars through the intersection per minute is 20, is the Poisson distribution appropriate?

SOLUTION:

(a) $\lambda = 3$ cars/30 seconds

we have $f(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$

$$P(X=0) = \frac{e^{-3}3^0}{0!} = 0.0498$$

(b) $P(X \geq 3) = 1 - P(X < 3) = 1 - [P(X=0) + P(X=1) + P(X=2)]$

$$= 1 - \left[\frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!} + \frac{e^{-3}3^2}{2!}\right] = 1 - [0.0498 + 0.1494 + 0.2240] = 0.5768$$

(c) $P(X \leq x) \geq 0.9$ or $x = 5$

(d) $\sigma^2 = \lambda = 6$. This is not appropriate.

Example E3.43

A student's campus newspaper in a particular university contains an average of 1.2 typographical errors per page.

(a) find the probability that a randomly selected page of this newspaper will contain exactly 4 typographical errors using the Poisson formula

(b) find the probability that the number of typographical errors on a randomly selected page will be

(i) more than 3

(ii) less than 4

Use the Poisson probabilities table in Appendix-*D*.

SOLUTION:

Let x be the number of typographical errors on a randomly selected page of this newspaper. Since it contains an average of 1.2 typographical errors per page, $\lambda = 1.2$. We have Eq. (3.62)

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

(a) $P(\text{exactly } 4) = P(x = 4) = \dfrac{\lambda^x e^{-\lambda}}{x!} = \dfrac{(1.2)^4 e^{-1.2}}{4!} = 0.0260$

(b) *(i)* $P(\text{more than } 3) = P(x > 3) = P(4) + P(5) + P(6) + P(7) = 0.0260 + 0.0062 + 0.0012 + 0.0002 = 0.0336$

(ii) $P(\text{less than } 4) = P(x < 4) = P(0) + P(1) + P(2) + P(3) = 0.3012 + 0.3614 + 0.2169 + 0.0867 = 0.9662$

Example E3.44

In a statistics course final examination, 15% of the students fail.

(a) find the probability that in a random sample of 100 students in that statistics class who took the final examination exactly 20 will fail. Use the Poisson formula.

(b) find the probability that the number of students who fail this statistics final examination in a randomly selected 100 students is

(i) at most 9

(ii) 10 to 16

(iii) at least 20

Use the Poisson probabilities table in Appendix-*D*.

SOLUTION:

Let x be the number of students in a random sample of 100 who fail the final examination in statistics. Since, an average, 15% of the students fail the examination, $\lambda = 0.15(100) = 15$.

(a) $P(\text{exactly 20 fail}) = P(x = 20) = \dfrac{e^{-\lambda}\lambda^x}{x!} = \dfrac{(e^{-15})(15)^{20}}{20!} = 0.0418$

(b) *(i)* $P(\text{at most 9 fail}) = P(x \le 9) = P(0) + P(1) + P(2) + P(3) + \cdots + P(9)$

$= 0.0000 + 0.0000 + 0.0000 + 0.0002 + 0.0006 + 0.0019 + 0.0048 + 0.0104 + 0.0194 + 0.0324$

$= 0.0697$

(ii) $P(\text{10 to 16 fail}) = P(10 \le x \le 16) = P(10) + P(11) + P(12) + P(13) + P(14) + P(15) + P(16)$

$= 0.0486 + 0.0663 + 0.0829 + 0.0956 + 0.1024 + 0.1024 + 0.0960 = 0.5942$

(iii) $P(\text{at least } 20) = P(x \ge 20) = P(20) + P(21) + P(22) + \cdots + P(39)$

$= 0.0418 + 0.0299 + 0.0204 + 0.0133 + 0.0083 + 0.0050 + 0.0029 + 0.0016 + 0.0009 + 0.0004$
$+ 0.0002 + 0.0001 + 0.0001 + 0.0000 + \cdots + 0.0000 = 0.1249$

Example E3.45

A hardware store in a big city receives an average of 9.8 telephone calls per hour.

(a) find the probability that exactly 6 telephone calls will be received at this store during a certain hour. Use the Poisson formula.

(b) find the probability that the number of telephone calls received at this store during a certain hour will be

(i) less than 8

(ii) more than 12

(iii) 5 to 8

Use the Poisson distribution table in Appendix-*D*.

SOLUTION:

Let x be the number of telephone calls received at this hardware store during a certain hour. Since the average number of telephone calls per hour is 9.8, $\lambda = 9.8$.

(a) $$P(\text{exactly } 6) = P(x = 6) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{(e^{-9.8})(9.8)^{6}}{6!} = 0.0677$$

(b) *(i)* $$P(\text{less than } 8) = P(x < 8) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + P(6) + P(7)$$
$$= 0.0001 + 0.0005 + 0.0027 + 0.0087 + 0.0213 + 0.0418 + 0.0682 + 0.0955 = 0.2388$$

(ii) $$P(\text{more than } 12) = P(x > 12) = P(13) + P(14) + \mathrm{P}(15) + \cdots$$
$$= 0.0685 + 0.0479 + 0.0313 + 0.0192 + 0.01111 + 0.0060 + 0.0031 + 0.0015 + 0.0007 + 0.0003$$
$$+ 0.0001 + 0.0001 = 0.1898$$

(iii) $$P(5 \text{ to } 8) = P(5 \le x \le 8) = P(5) + P(6) + P(7) + P(8) = 0.0418 + 0.0682 + 0.0955 + 0.1170 = 0.3225$$

Example E3.46

An average of 0.7 accidents occur per day in a large city.

(a) find the probability that no accidents will occur in that city on a given day

(b) write the probability distribution of x, where x denotes the number of accidents that will occur in that city on a given day

(c) find the mean, variance and standard deviation of the probability distribution developed in part *(b)*.

SOLUTION:

x = number of accidents on a given day in that city.

Since the average number of accidents per day is 0.7, $\lambda = 0.7$.

(a) $$P(\text{no accidents}) = P(x = 0) = \frac{e^{-\lambda}\lambda^{x}}{x!} = \frac{e^{-0.7}(0.7)^{0}}{0!} = 0.4966$$

(b) From the table in Appendix-*D*, we have

x	P(x)
0	0.4966
1	0.3476
2	0.1217
3	0.0284
4	0.0050
5	0.0007
6	0.0001
7	0.0000

(*c*) The mean, variance and standard deviation are:

mean $\mu = \lambda = 1.7$

variance $\sigma^2 = \lambda = 0.7$ and

standard deviation, $\sigma = \sqrt{\lambda} = \sqrt{0.7} = 0.8366$

Example E3.47

The number of male mates of a queen bee was found to have a Poisson distribution with parameter $\lambda = 2.7$. Find the probability that the number, *N*, of male mates of a queen bee is

(*a*) exactly 2
(*b*) at most 2
(*c*) between 1 and 3, inclusive
(*d*) on average, how many male mates does a queen bee have?
(*e*) develop a table of probabilities for the random variable, *N*. Compute the probabilities until they are zero to 4 decimal places
(*f*) draw a histogram of the probabilities in part (*c*).

SOLUTION:

(*a*) $P(N = 2) = e^{-2.7}\dfrac{2.7^2}{2!} = 0.2450$

(*b*) $P(N \le 2) = P(0 \text{ or } 1 \text{ or } 2) = e^{-2.7}\dfrac{2.7^0}{0!} + e^{-2.7}\dfrac{2.7^1}{1!} + e^{-2.7}\dfrac{2.7^2}{2!} = 0.0672 + 0.1815 + 0.2450 = 0.4937$

(*c*) $P(1 \le N \le 3) = P(1 \text{ or } 2 \text{ or } 3) = e^{-2.7}\dfrac{2.7^1}{1!} + e^{-2.7}\dfrac{2.7^2}{2!} + e^{-2.7}\dfrac{2.7^3}{3!} = 0.1815 + 0.2450 + 0.2205 = 0.6470$

(*d*) On average, the number of mates of a queen bee is

$\mu = \lambda = 2.7$

(*e*) Refer to Appendix-*D*. The results are shown in Table E3.47.

Table E3.47

N	P(N = n)
0	0.0672
1	0.1815
2	0.2450
3	0.2205
4	0.1488
5	0.0804
6	0.0362
7	0.0139
8	0.0047
9	0.0014
10	0.0004
11	0.0001
12	0.0000

(*f*) See Fig. E3.47

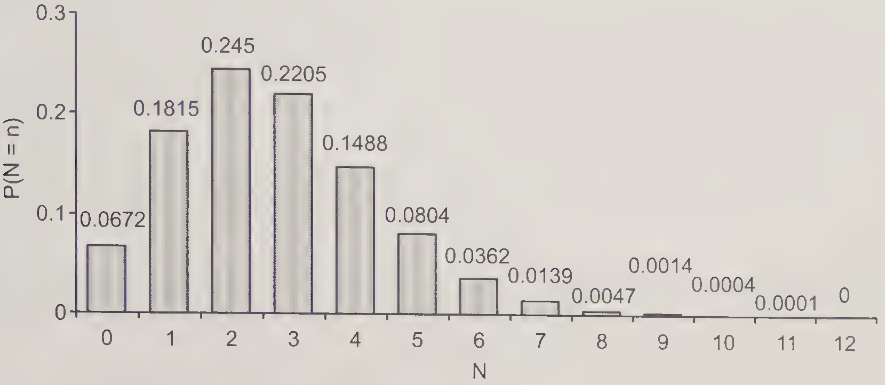


Fig. E3.47

3.4 CONTINUOUS PROBABILITY DISTRIBUTIONS

3.4.1 The Normal Distribution

In everyday life, people deal with and use a wide variety of variables. Some of these variables — such as heights of people, scores in final examinations, aptitude test scores, TOEFL (Test of English as a Foreign Language), and GRE (Graduate Record Examination) share an important characteristic: their distributions have roughly the shape of a normal curve, that is, a special type of bell-shaped curve like one shown in Fig. 3.6.

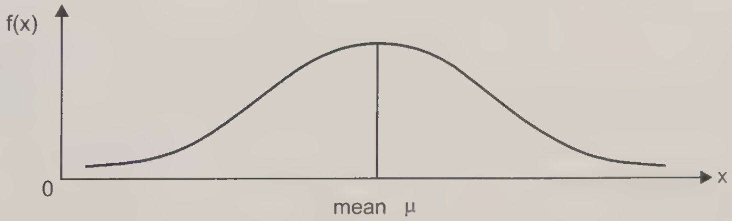


Fig. 3.6 Normal curve

The normal distribution is in many respects the cornerstone of statistics. A random variable X is said to have a normal distribution with mean $\mu(-\infty < \mu < \infty)$ and variance $\sigma^2 > 0$ if it has the density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)\sigma]^2} \qquad -\infty < X < \infty \tag{3.71}$$

The distribution is illustrated graphically in Fig. 3.6. The normal distribution is used so extensively that the shorthand notation $X \sim N(\mu, \sigma^2)$ is often used to indicate that the random variable X is normally distributed with mean μ and variance σ^2.

A variable is said to be a normally distributed variable or to have a *normal distribution* if its distribution has the shape of a normal curve.

If a variable of a population is normally distributed and is the only variable under consideration, the general practice is to assume that the population is *normally distributed* or that it is a *normally distributed population*. In reality, a distribution is unlikely to have exactly the shape of a normal distribution curve. If a variable's distribution is shaped roughly like a normal curve, we can consider that the variable is an *approximately normally distributed variable* or that it has *approximately a normal distribution.*

A normal distribution (and hence a normal curve) is completely determined by the mean and standard deviation. Hence, two normally distributed variables having the same mean and standard deviation must have the same distribution. We identify a normal curve by stating the corresponding mean and standard deviation and calling those to *parameters* of the normal curve.

A normal distribution is symmetric about and centred at the mean of the variable, and its spread depends on the standard deviation of the variable. The larger the standard deviation, the flatter and more spread out is the distribution.

In summary, the normal curve associated with a normal distribution is bell shaped, centered at μ, and close to the horizontal axis outside the range from $\mu - 3\sigma$ to $\mu + 3\sigma$ as shown in Figs. 3.6 and 3.6(*a*).

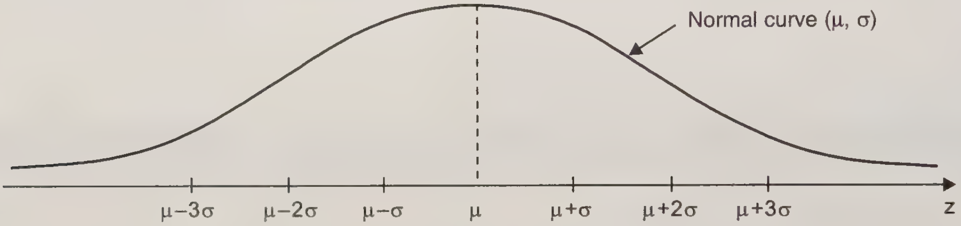


Fig. 3.6(a)

Random variables with different means and variances can be modeled by normal probability functions with appropriate choices of the centre and width of the curve. The value of $E(X) = \mu$ determines the centre of the probability density function and the value of $V(X) = \sigma^2$ determines the width. Figure 3.6(*c*) illustrates several normal probability density functions with selected values of m and σ^2. Each one has the characteristic symmetric bell-shaped curve, but the centers and dispersions differ.

Definition of Normal Distribution

A random variable X with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\left(\frac{x-\mu}{2\sigma^2}\right)} \qquad -\infty < x < \infty$$

is a *normal random variable* with parameters μ, where $-\infty < x < \infty$, and $\sigma > 0$.

Also $\quad E(X) = \mu$ and $V(X) = \sigma^2$

and the notation $N(\mu, \sigma^2)$ is used to denote the distribution. The mean and variance of X are equal to μ and σ^2, respectively.

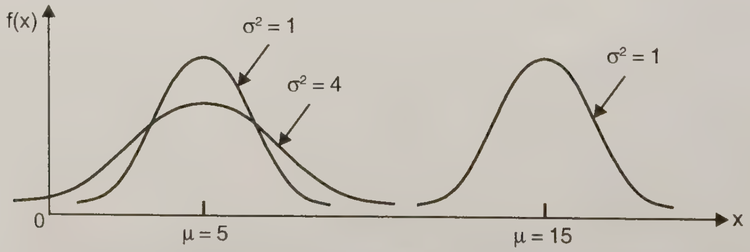


Fig. 3.6(b): Normal probability density functions for selected values of the parameters μ and σ²

3.4.1.1 Properties of the Normal Distribution

The normal distribution has several important properties:

1. $\int_{\infty}^{\infty} f(x)\,dx = 1$
2. $f(x) \geq 0$ for all x
3. $\lim_{x \to \infty} f(x) = 0$ and $\lim_{x \to \infty} f(x) = 0$
4. $f[(x + \mu)] = f[-(x - \mu)]$. The density is symmetric about μ.
5. The maximum value of $f(x)$ occurs at $x = \mu$.
6. The points of inflection of $f(x)$ are at $x = \mu \pm \sigma$.

Also, $P(a \leq x \leq b) = \int_{a}^{b} f(x)\,dx$ = area under $f(x)$ from a to b for any a and b.

Property 1 may be demonstrated as follows. Let $y = (x - \mu)/\sigma$ in Eq. (3.71) and denote the integral as I. That is,

$$I = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} e^{-(1/2)y^2}\,dy \frac{1}{\sqrt{2\pi}} \int_{\infty}^{\infty} e^{-(1/2)x^2}\,dx \tag{3.72}$$

on changing to polar coordinates with the transformation of variables $y = r \sin \theta$ and $z = r \cos \theta$, the integral becomes

$$I^2 = \frac{1}{2\pi} \int_{0}^{\infty} \int_{0}^{2\pi} re^{-(1/2)r^2}\,d\theta\,dr = \int_{0}^{\infty} re^{-(1/2)r^2}\,dr = 1 \tag{3.73}$$

3.4.1.2 Mean and Variance of the Normal Distribution

The mean of the normal distribution may be determined easily. Since

$$E(X) = \int_{-\infty}^{\infty} \frac{x}{\sigma\sqrt{2\pi}} e^{-(1/2)[(x-\mu)/\sigma^2]}\,dx \tag{3.74}$$

and if we let $z = (x - \mu)/\sigma$, we obtain

$$E(X) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} (\mu + \sigma z) e^{-z^2/2}\,dz \tag{3.75}$$

$$= \mu = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\,dz + \sigma \frac{1}{\sqrt{2\pi}} z e^{-z^2/2}\,dz \tag{3.76}$$

Since the integrand of the first integral is that of a normal density with $\mu = 0$ and $\sigma^2 = 1$, the value of the first integral is one. The second integral has value zero, that is,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z e^{-z^2/2}\,dz = \frac{1}{\sqrt{2\pi}} z e^{-z^2/2} \Big|_{-\infty}^{\infty} = 0$$

and thus

$$E(X) = \mu[1] = \sigma[0] = \mu \tag{3.77}$$

To find the variance we must evaluate

$$V(X)=E[(X-\mu)^2]=\int_{-\infty}^{\infty}(x-\mu)^2\frac{1}{\sigma\sqrt{2\pi}}e^{-(1/2)[(x-\mu)/\sigma]^2}dx$$

And letting $z=(x-\mu)/\sigma$, we obtain

$$V(X)=\int_{-\infty}^{\infty}\sigma^2 z^2\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz=\sigma^2\left[\int_{-\infty}^{\infty}\frac{z^2}{\sqrt{2\pi}}e^{-z^2/2}dz\right]$$

$$=\sigma^2\left[\left.\frac{-ze^{-z^2/2}}{\sqrt{2\pi}}\right|_{-\infty}^{\infty}\right]+\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz=\sigma^2(0+1)$$

so that $$V(X)=\sigma^2 \tag{3.78}$$

In summary the mean and variance of the normal density given in Eq. (3.78) are μ and σ^2, respectively.

3.4.1.3 The Cumulative Normal Distribution

The distribution function F is

$$F(x)=P(X\le x)=\int_{-\infty}^{\infty}\frac{1}{\sigma\sqrt{2\pi}}e^{-(1/2)[(x-\mu)/\sigma]^2}du \tag{3.79}$$

It is impossible to evaluate this integral without resorting to numerical methods, and even then the evaluation would have to be accomplished for each pair (μ, σ^2). However, a simple transformation of variables, $z=(X-\mu)/\sigma$, allows the evaluation to be independent of μ and σ.

That is,

$$F(X)=P(X\le x)=P\left(Z\le\frac{x-\mu}{\sigma}\right)=\int_{-\infty}^{(x-\mu)/\sigma}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz=\int_{-\infty}^{(x-\mu)/\sigma}\varphi(z)\,dz=\phi\left(\frac{x-\mu}{\sigma}\right) \tag{3.80}$$

3.4.1.4 The Standard Normal Distribution

The probability distribution in Eq. (3.80) above,

$$\varphi(z)=\frac{1}{\sqrt{2\pi}}e^{-z^2/2} \qquad -\infty<x<\infty$$

is a normal distribution with mean 0 and variance 1; that is, $Z \sim N(0, 1)$ and we say that Z has a *standard normal distribution.* A graph of the probability density function is shown in Fig. 3.7.

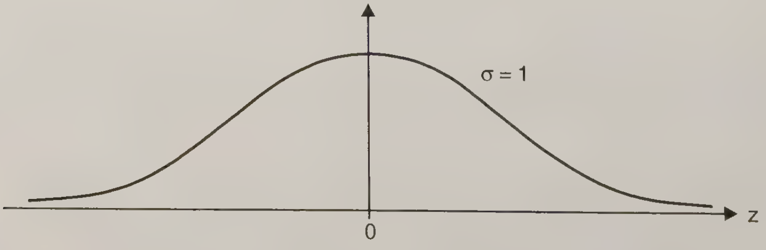


Fig. 3.7: The standard normal distribution

The corresponding distribution function is ϕ, where

$$\varphi(z) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} du \tag{3.81}$$

where
$$z = \frac{x-\mu}{\sigma}$$

and this function has been well tabulated.

A few useful results for a normal distribution are summarised below and shown in Fig. 3.7(*a*). For random variable,

$$P(\mu - \sigma < X < \mu + \sigma) = 0.6827$$
$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$$
$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$$

In other words, any normally distributed variable has the following properties:

1. 68.27% of all possible observations lie within one standard deviation to either side of the mean, that is, between $\mu - \sigma$ and $\mu + \sigma$.
2. 95.45% of all possible observations lie within two standard deviation to either side of the mean, that is, between $\mu - 2\sigma$ and $\mu + 2\sigma$.
3. 99.73% of all possible observations lie within three standard deviation to either side of the mean, that is, between $\mu - 3\sigma$ and $\mu + 3\sigma$.

These properties are shown in Fig. 3.7(*b*).

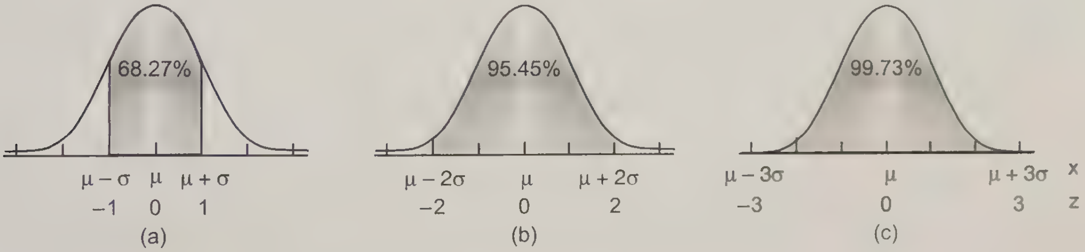


Fig. 3.7(a)

In addition, from the symmetry of $f(x)$, $P(X > \mu) = P(X < \mu) = 0.5$. The above model assigns some probability to each interval of the real line since $f(x)$ is positive for all x. The probability density function decreases as x moves further from μ. As a result, the probability that a measurement falls far from μ is small, and at some distance from μ the probability of an interval can be approximated as zero.

For a normally distributed variable, the percentage of all possible observations that (i.e., within any specified range) equals the corresponding area under its associated normal curve, expressed as a percentage. This result holds approximately for a variable that is approximately normally distributed.

A normally distributed variable having mean 0 and standard deviation 1 is said to have the *standard normal distribution*. Its associated normal curve is called the *standard normal curve*, which is shown in Fig. 3.7(*c*).

The area under a normal probability density function beyond 3σ from the mean is quite small. Since more than 0.9973 of the probability of a normal distribution is within the interval $(\mu - 3\sigma, \mu + 3\sigma)$, 6σ is often referred to as the *width* of a normal distribution. Advanced integration methods can be used to show that the area under the normal probability density function from $-\infty < x < \infty$ is 1.

3.4.1.5 Problem-Solving Procedure

If X is a normal random variable with $E(X) = \mu$ and $V(X) = \sigma^2$, the random variable

$$Z = \frac{X - \mu}{\sigma}$$

is a normal random variable with $E(Z) = 0$ and $V(Z) = 1$. That is, Z is a standard normal random variable.

Suppose X is a normal random variable with mean μ and variable σ^2. Then

$$P(X \le x) = P\left(\frac{X - \mu}{\sigma} \le \frac{x - \mu}{\sigma}\right) = P(Z \le z)$$

where Z is a *standard normal variable*, and $z = \left(\dfrac{x - \mu}{\sigma}\right)$ is the z-value obtained by standardising X.

The probability is obtained by using the table in Appendix-*E* with $z = (x - \mu)/\sigma$.

Basic Properties of the Standard Normal Curve

1. The total area under the standard normal curve is 1.
2. The standard normal curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.
3. The standard normal curve is symmetric about 0. That is, the part of the curve to the left of the vertical line at the centre in Fig. 3.7(*b*) is the mirror image of the curve to the right of it.
4. Almost all the area under the standard normal curve lies between –3 and 3.

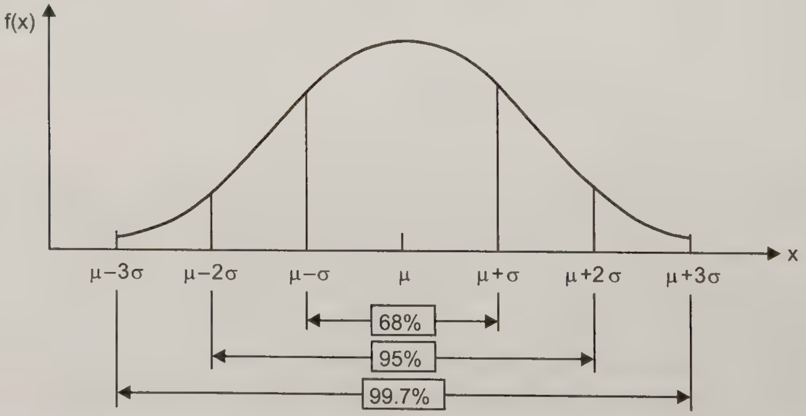


Fig. 3.7(b): Probabilities associated with a normal distribution

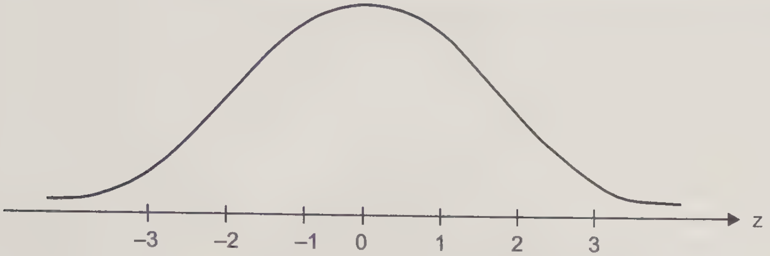


Fig. 3.7(c): Standard normal distribution

A normal random variable with $\mu = 0$ and $\sigma^2 = 1$ is called a *standard normal variable* and is denoted as Z. The cumulative distribution function of a standard normal random variable is denoted as $\phi(z) = P(Z \le z)$.

The tables in Appendix-*E* provides cumulative probabilities for a standard normal random variable.

The procedure in solving problems involving calculating the cumulative normal probabilities is simple. For example, suppose that $X \sim N(50, 4)$, and we wish to find the probability that x is less than or equal to 54; that is $P(x \le 54) = F(54)$. Since that standard normal random variable is

$$z = \frac{x-\mu}{\sigma}$$

we can *standardize* the point of interest $x = 104$ to obtain

$$z = \frac{x-\mu}{\sigma} = \frac{54-50}{2} = 2$$

Now the probability that the *standard* normal random variable z is less than or equal to 2 is equal to the probability that the *original* normal random variable x is less than or equal to 54. Expressed mathematically,

$$F(x) = \phi\left(\frac{x-\mu}{\sigma}\right) = \phi(z)$$

$$F(54) = \phi(2)$$

The table in Appendix-*E* contains cumulative standard normal probabilities for various values of z. From this table, we can read

$$\phi(2) = 0.9772$$

Note that in the relationship $z = (x - \mu)/\sigma$, the variable z measures the departure of x from the mean μ in standard deviation (σ) units. In our example $F(54) = f(2)$ indicates that 54 is *two* standard deviations ($\sigma = 2$) above the mean. In general, $x = \mu + \sigma z$. In solving problems, we sometimes need to use the symmetry property of φ in addition to the tables.

In order to find the percentages of all possible observations of a normally distributed variable that lie within any specified range, we express the range in terms of z-scores and then determine the corresponding area under the standard normal curve. The stepwise procedure to determine a percentage or probability for a normally distributed variable is presented below:

1. Sketch the normal curve associated with the variable as shown in Fig. 3.7(*d*).
2. Shade the region of interest and mark its delimiting x-value(s).

3. Compute the z-scores for the delimiting x-value(s) found in step 2.
4. Use the table in Appendix-E to find the area under the standard normal curve delimited by the z-score(s) found in step 3.

The probability distributions presented in this chapter are summarised in Table 3.1.

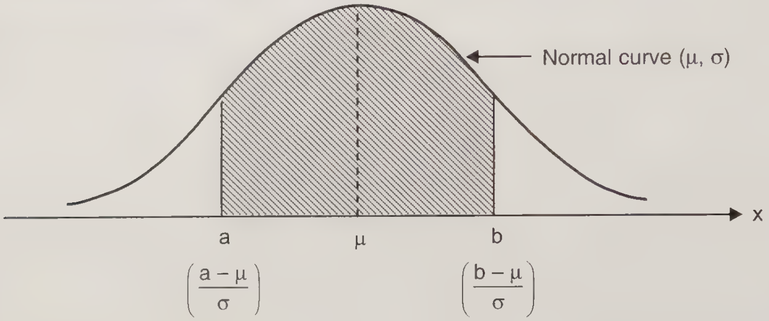

Normal curve (μ, σ)
x
a
μ
b
$\left(\frac{a-\mu}{\sigma}\right)$
$\left(\frac{b-\mu}{\sigma}\right)$

Fig. 3.7(d)

Table 3.1: Summary of probability distributions

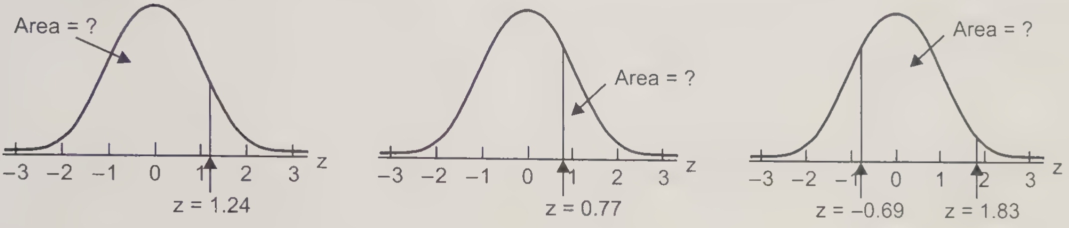
Name	Probability Density Function	Mean	Variance
Discrete			
Hypergeometric	$\dfrac{\binom{K}{x}\binom{N-k}{n-x}}{\binom{N}{n}}$ $x = \max(0, n - N + K), 1, \ldots \min(K, n), K \leq N, n \leq N$	np, where $p = \dfrac{K}{N}$	$np(1-p)\left(\dfrac{N-m}{N-1}\right)$
Binomial	$\binom{n}{x} p^x (1-p)^{n-x}$, $x = 0, 1, \ldots, n, 0 \leq p \leq 1$	np	$np(1-p)$
Poisson	$\dfrac{e^{-\lambda}\lambda^x}{x!}, x = 0, 2, \ldots, \lambda > 0$	λ	λ
Continuous			
Normal	$\dfrac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ $-\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$	μ	σ^2

Example E3.48

(a) Determine the area under the standard normal curve that lies to the left of 1.24, as shown in Fig. E3.48(a).

(*b*) Determine the area under the standard normal curve that lies to the right of 0.77, as shown in Fig. E3.48(*b*).

(*c*) Determine the area under the standard normal curve that lies between –0.69 and 1.83, as shown in Fig. E3.48(*c*).



(a) (b) (c)

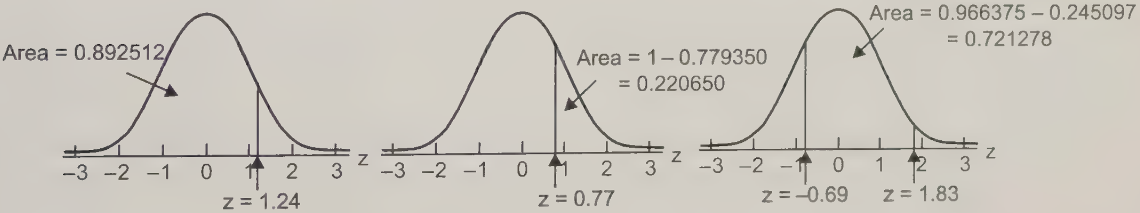
Fig. E3.48

SOLUTION:

(*a*) Refer to the table in Appendix-*E*. First, we go down the left hand column, labeled z, to "1.2". Then, going across that row to the column labeled "0.04", we reach 0.892512. This number is the area under the standard normal curve that lies to the left of 1.24, as shown in Fig. E3.48(*d*).

(*b*) Because the total area under the standard normal curve is 1, the area to the right of 0.77 equals 1 minus the area of the left of 0.77. We find this by first giving down the z-column to "0.7". Then, going across that row to the column, labeled "0.07", we reach 0.779350, which is the area under the standard normal curve that lies to the left of 0.77. Thus, the area under the standard normal curve that lies to the right of 0.77 is 1 – 0.779350 = 0.220650 as shown in Fig. E3.48(*e*).

(*c*) The area under the standard normal curve that lies between –0.69 and 1.83 equals the area to the left of 1.83 minus the area to the left of –0.69. The table in Appendix-*E* shows that these latter two areas are 0.966375 and 0.245097, respectively. Hence the area we seek is 0.966375 – 0.245097 = 0.721278, as shown in Fig. E3.48(*f*).



(d) (e) (f)

Fig. E3.48

Example E3.49

Determine the *z*-score having an area of 0.06 to its left under the standard normal curve, as shown in Fig. E3.49.

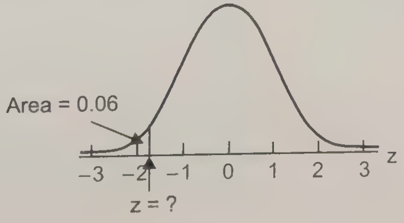


Fig. E3.49

SOLUTION:

See Fig. E3.48(*a*).

Search the body of the table in Appendix-*E* for the area 0.06. There is no such area, so use the area closest to 0.06, which is 0.060571. The z-score corresponding to that area is –1.55. Hence, the z-score having area 0.06 to its left under the standard normal curve is roughly –1.55, as shown in Fig. E3.49(*a*).

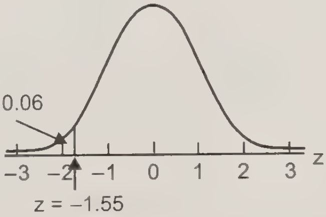


Fig. E3.49(a)

Example E3.50

Use the table in Appendix-*E* to find

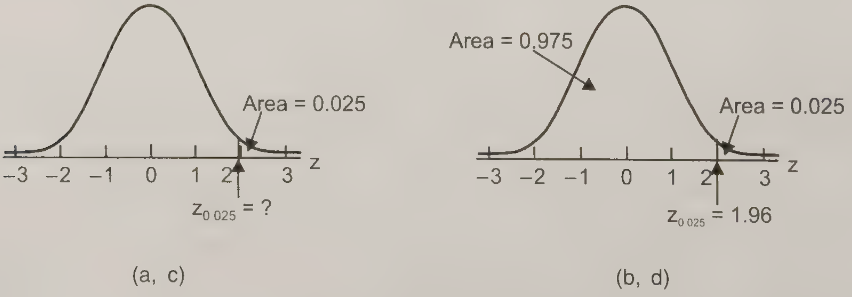
(*a*) $z_{0.025}$

(*b*) $z_{0.05}$

SOLUTION:

(*a*) $z_{0.025}$ is the z-score that has an area of 0.025 to its right under the standard normal curve, as shown in Fig. E3.50(*a*). Because the area to its right is 0.025, the area to its left is 1 – 0.025 = 0.975, as shown in Fig. E3.50(*b*). Table in Appendix-*E* contains an entry for the area 0.975002, its corresponding z-score is 1.96. Thus, $z_{0.025} = 1.96$, as shown in Fig. E3.50(*b*).

(*b*) $z_{0.05}$ is the z-score that has an area of 0.05 to its right under the standard normal curve, as shown in Fig. E3.50(*c*). Because the area to its right is 0.05, the area to its left is 1 – 0.05 = 0.95, as shown in Fig. E3.50(*d*). Table in Appendix-*E* does not contain any entry for the area 0.95 and has two area entries equally closest to 0.95 – namely, 0.949497 and 0.950529. The z-scores corresponding to those two areas are 1.64 and 1.65 respectively. So our approximation of $z_{0.05}$ is the mean of 1.64 and 1.65, that is, $z_{0.05} = 1.645$, as shown in Fig. E3.50(*d*).



(a, c) (b, d)

Fig. E3.50

Example E3.51

Find the two z-scores that divide the area under the standard normal curve into a middle 0.95 area and two outside 0.025 areas, as shown in Fig. E3.51(*a*).

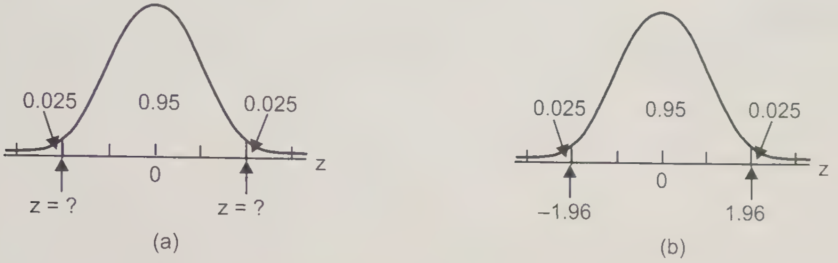


Fig. E3.51

SOLUTION:

The area of the shaded region on the right in Fig. E3.51(*a*) is 0.025. The corresponding z-score, $z_{0.025}$, is 1.96. Because the standard normal curve is symmetric about 0, the z-score on the left is −1.96. Therefore the two required z-scores are ±1.96, as shown in Fig. E3.51(*b*).

Example E3.52

In a long distance run of 10 km in a city the times of finishers are normally distributed with mean 71 minutes and standard deviation 9 minutes.

(*a*) determine the percentage of finishers with times between 60 and 80 minutes

(*b*) find the percentage of finishers with times less than 85 minutes

(*c*) obtain and interpret the 40th percentile for the finishing times

(*d*) find and interpret the 8th decile for the finishing times.

SOLUTION:

(*a*) For finishers with times of 60 and 80 minutes, the z-values are

$$z = \frac{x-\mu}{\sigma} = \frac{60-71}{9} = -1.22 \text{ and } z = \frac{x-\mu}{\sigma} = \frac{80-71}{9} = 1.00$$

The area to the left of $z = -1.225$ is 0.111233 and the area to the left of $z = 1.00$ is 0.841345. Hence, the area between $z = -1.22$ and $z = 1.00$ is $0.841345 - 0.111233 = 0.730112$. Thus, the percentage of finishers with times between 60 minutes and 80 minutes in that city 10 km run is 73.01%.

(*b*) For finishing time of 85 minutes, the z-value is

$$z = \frac{x-\mu}{\sigma} = \frac{85-71}{9} = 1.56$$

The area of the left of $z = 1.56$ is 0.9406. Thus, the percentage of finishers with times less than 85 minutes is 94.06%.

(*c*) Using the table in Appendix-*E*, we find that an area of 0.40 lies to the left of $z = -0.25$. We convert this z-value to an x-value using $x = \mu + z\sigma$. Hence 40% of the finishers had times less than 40th percentile, $x = 71 + (-0.25)(9) = 68.75$ minutes.

(*d*) The 8th decile is the same as the 80th percentile. Using the table in Appendix-*E*, we find that an area of 0.80 lies to the left of $z = 0.84$. We convert the z-value to an x-value using $x = \mu + z\sigma$. Thus, 80% of the finishing times were less than $71 + (0.84)9 = 78.56$ minutes.

Example E3.53

If X has a normal distribution for which the mean is 1 and standard deviation is 2, find the value of each of the following:

(*a*) $P(X \le 3)$

(*b*) $P(X > 1.5)$

(*c*) $P(X = 1)$

(*d*) $P(2 < X < 5)$

(*e*) $P(X \ge 0)$

(*f*) $P(-1 < X < 0.5)$

(*g*) $P(|X| \le 2)$

(*h*) $P(1 \le -2X + 3 \le 8)$

SOLUTION:

Given $\mu = 1$ and standard deviation = 2. Refer to the table in Appendix-*E* for cumulative standard normal standard distribution.

$$z = \frac{x - \mu}{\sigma} = \frac{x - 1}{2}$$

then z has a standard normal distribution.

(*a*) $P(X \le 3) = P(Z \le 1) = \Phi(1) = 0.841345$

(*b*) $P(X > 1.5) = P(Z > 0.25) = 1 - \Phi(0.25) = 0.4013$

(*c*) $P(X = 1) = 0$, because X has a continuous distribution.

(*d*) $P(2 < X < 5) = P(0.5 < Z < 2) = \Phi(2) - \Phi(0.5) = 0.2858$

(*e*) $P(X \ge 0) = P(Z \ge -0.5) = P(Z \le 0.5) = \Phi(0.5) = 0.6915$

(*f*) $P(-1 < X < 0.5) = P(-1 < Z < -0.25) = P(0.25 < Z < 1) = \Phi(1) - \Phi(0.25) = 0.2426$

(*g*) $P(|X| \le 2) = P(-2 \le X \le 2) = P(-1.5 \le Z \le 0.5) = P(Z \le 0.5) - P(Z \le -1.5) = P(Z \le 0.5) - P(Z \le 1.5)$
$= \Phi(0.5) - [1 - \Phi(1.5)] = 0.6247$

(*h*) $P(1 \le -2X + 3 \le 8) = P(-2 \le -2X \le 5) = P(-2.5 \le X \le 1) = P(-1.75 \le Z \le 0) = P(0 \le Z \le 1.75)$
$= \Phi(1.75) - \Phi(0) = 0.4599$

Example E3.54

Assume X is normally distributed with a mean of 10 and a standard deviation of 2. Determine the following:

(*a*) $P(X < 13)$

(*b*) $P(X > 9)$

(*c*) $P(6 < X < 14)$

(*d*) $P(2 < X < 4)$

(*e*) $P(-2 < X < 8)$

SOLUTION:

Refer to the table in Appendix-*E*.

(*a*) $P(X < 13) = P\left(Z < \dfrac{13-10}{2}\right) = P(Z < 1.5) = 0.93319$

(*b*) $P(X > 9) = 1 - P(X < 9) = 1 - P(Z < -0.5) = 0.69146$

(*c*) $P(6 < X < 14) = P\left(\dfrac{6-10}{2} < Z < \dfrac{14-10}{2}\right) = P(-2 < Z < 2) = [P(Z < 2) - P(Z < -2)] = 0.9545$

(*d*) $P(2 < X < 4) = P\left[\dfrac{2-10}{2} < Z < \dfrac{4-10}{2}\right] = P(-4 < Z < -3) = P(Z < -3) - P(Z < -4) = 0.00132$

(*e*) $P(-2 < X < 8) = P(X < 8) - P(X < -2) = P\left(Z < \dfrac{8-10}{2}\right) - P\left(Z < \dfrac{-2-10}{2}\right)$

$= P(Z < -1) - P(Z < -6) = 0.15866$

Example E3.55

The compressive strength of samples of cement can be modeled by a normal distribution with a mean of 7000 kg/cm^2 and a standard deviation of 100 kg/cm^2.

(*a*) find the probability of a sample's strength that is less than 7250 kg/cm^2

(*b*) find the probability that a sample's strength is between 6800 and 6900 kg/cm^2

(*c*) what strength is exceeded by 95% of the samples?

SOLUTION:

Refer to the table in Appendix-*E*.

(*a*) $P(X < 7250) = P\left(Z < \dfrac{7250-7000}{100}\right) = P(Z < 2.5) = 0.99379$

(*b*) $P(6800 < X < 6900) = P\left[\dfrac{6800-7000}{100} < Z < \dfrac{6900-7000}{100}\right] = P(-2 < Z < -1)$

$= P(Z < -1) - P(Z < -2) = 0.13591$

(*c*) $P(X > x) = P\left(Z > \dfrac{x-7000}{100}\right) = 0.95$

Hence, $\dfrac{x-7000}{100} = -1.65$ and $x = 6835$.

Example E3.56

(*a*) A process manufactures ball bearings whose diameters are normally distributed with mean 3.505 cm and standard deviation 0.008 cm. Specifications call for the diameter to be in the interval 3.5 ± 0.01 cm. What proportion of the ball bearings will meet the specification?

(*b*) Suppose the process can be recalibrated so that the mean will equal to 3.5 cm, the centre of the specification interval. The standard deviation of the process remains 0.008 cm. What proportion of the diameter will meet the specifications?

SOLUTION:

Let X represents the diameter of a randomly chosen ball bearing. Then, $X \sim N(3.505, 0.008^2)$. Figure E3.56(*a*) shows the probability density function of the $N(3.505, 0.008^2)$ population. The shaded area represents $P(3.49 < X < 3.51)$, which is the proportion of ball bearings that meet the specification.

(*a*) We compute the z-scores of 3.49 and 3.51 as follows:

$$z = \frac{3.49 - 3.505}{0.008} = -1.88 \text{ and } z = \frac{3.51 - 3.505}{0.008} = 0.63$$

The area to the left of $z = -1.88$ is 0.0301. The area to the left of $z = 0.63$ is 0.7357. The area between $z = 0.63$ and $z = -1.88$ is $0.7357 - 0.0301 = 0.7056$. Hence, approximately 70.56% of the diameters of the ball bearings will meet the specifications.

(*b*) The mean is 3.5000 rather than 3.505. The calculations are as follows: (see Fig. E3.56(*b*)).

$$z = \frac{3.49 - 3.50}{0.008} = -1.25 \text{ and } z = \frac{3.51 - 3.50}{0.008} = 1.25$$

The area to the left of $z = -1.25$ is 0.1056. The area of the left of $z = 1.25$ is 0.8944. The area between $z = 1.25$ and $z = -1.25$ is $0.8944 - 0.1056 = 0.7888$. Hence, recalibrating will increase the proportion of diameters that meet the specification to 78.88%.

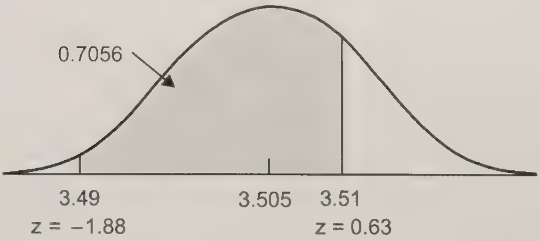


Fig. E3.56(a)

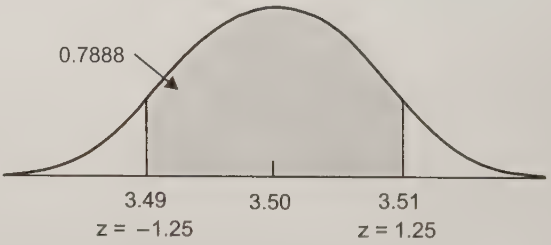


Fig. E3.56(b)

Example E3.57

Assume X is normally distributed with a mean of 5 and a standard deviation of 1. Determine the value of x that solves each of the following:

(*a*) $P(X > x) = 0.5$

(*b*) $P(X > x) = 0.95$

(*c*) $P(x < X < 5) = 0.2$

(*d*) $P[-x < (X - 5) < x] = 0.95$

(*e*) $P[-x < (X - 5) < x] = 0.99$

SOLUTION:

(*a*) $P(X < x) = P\left(Z > \dfrac{x-5}{1}\right) = 0.5$

Therefore, $\dfrac{x-5}{1} = 0$ and $x = 5$.

(*b*) $P(X < x) = P\left(Z > \dfrac{x-5}{1}\right) = 1 - P\left(Z < \dfrac{x-5}{1}\right) = 0.95$

Therefore, $P\left(Z < \dfrac{x-5}{1}\right) = 0.05$ and $\dfrac{x-5}{2} = -1.64$ and hence $x = 3.36$.

(*c*) $P(x < X < 5) = P\left(\dfrac{x-5}{1} < Z < 0\right) = 0.2$

Therefore, $P\left(Z < \dfrac{x-5}{1}\right) = 0.3$ and $\dfrac{x-5}{2} = -0.52$ and hence $x = 4.48$

(*d*) $P[5 - x < X < 5 + x] = P\left(\dfrac{-x}{2} < X < \dfrac{x}{2}\right) = 0.95$

Therefore, $x/2 = 0.98$

and $x = 1.96$

(*e*) $P[5 - x < X < 5 + x] = P[-x/2 < X < x/2] = 0.99$

Therefore, $x/2 = 1.79$

and $x = 2.58$.

Example E3.58

The diameter of a metal shaft for a precision instrument is assumed to be normally distributed with a mean of 0.5 mm and a standard deviation of 0.025 mm.

(*a*) what is the probability that shaft diameter is greater than 0.31 mm?

(*b*) what is the probability that shaft diameter is between 0.235 and 0.315 mm?

(*c*) the diameter of 90% of samples is below what value?

SOLUTION:

(*a*) $P(X > 0.31) = P\left(Z > \dfrac{0.315 - 0.25}{0.025}\right) = P(Z > 3) = 0.00135.$

(*b*) $P(0.235 < X < 0.315) = P(-0.6 < Z < 2.6) = P(Z < 2.6) - P(Z < -0.6) = 0.99534 - 0.27425 = 0.72109.$

(*c*) $P(X < x) = P\left(Z < \dfrac{x - 0.25}{0.025}\right) = 0.90$

Hence, $\dfrac{x - 0.25}{0.025} = 1.28$

and $x = 0.282$

Example E3.59

The length of a metal rod used in a machine system is normally distributed with a length of 45.10 mm and a standard deviation of 0.05 mm.

- *(a)* what is the probability that a rod is longer than 45.15 mm or shorter than 49.85 mm?
- *(b)* what should the process mean be set as to obtain the greatest number of rods between 44.85 and 45.15 mm?
- *(c)* if the rods that are not between 44.85 and 45.15 mm are scrapped, what is the yield for the process mean that one would select in part *(b)*?

SOLUTION:

(a) $$P(45.15 < X) = P\left[\left(\frac{45.15-45.10}{0.05}\right) < Z\right] = P(1 < Z) = 1 - 0.084134 = 0.15866$$

$$P(X < 49.85) = P\left(Z < \frac{49.85-45.10}{0.05}\right) = P(Z < -5) = 0$$

Hence, the answer is 0.159.

(b) The process mean should set at the centre of the specifications, that is, at $\mu = 45.0$

(c) $$P(44.85 < X < 45.15) = P\left[\frac{44.85-45.0}{0.05}\right] < Z < \left[\frac{45.15-45.0}{0.05}\right] = P(-3 < Z < 3) = 0.9973$$

Example E3.60

Refer to Example E3.59. If the process is centred so that the mean is 45 mm and the standard deviation is 0.05 mm and that 10 rods are measured and are assumed to be independent, determine

- *(a)* the probability that all 10 rods measured are between 45.85 and 45.15 mm
- *(b)* the expected number of the 10 rods that are between 44.85 and 45.15 mm.

SOLUTION:

(a) $P(44.85 < X < 45.15) = 0.9973$ from Example E3.59. Therefore, by independence, the probability of 10 rods are within the given limit is $0.9973^{10} = 0.9733$.

(b) Let Y denotes the number of rods from the 10 selected that are within the given limits. Then, Y is binomial with $n = 10$ and $p = 0.9973$.

Hence, $E(Y) = 9.973$.

Example E3.61

The weight of a electronic component is normally distributed with a mean of 6 ounces and a standard deviation of 0.25 ounce.

- *(a)* find the probability that the electronic component weighs more than 6.5 ounces.
- *(b)* what must the standard deviation of weight be in order for the company to state that 99% of its electronic components are less than 6.5 ounces?

(*c*) if the standard deviation stays at 0.25 ounces, what must the mean weight be in order for the company to state that 99.9% of its electronic components are less than 6.5 ounces?

SOLUTION:

(*a*) $P(X > 6.5) = P\left(Z > \dfrac{6.5-6}{0.25}\right) = P(Z > 2) = 0.02275$

(*b*) If $P(X < 13) = 0.999$, then $P\left(Z < \dfrac{6.5-6}{\sigma}\right) = 0.999$

Hence, $\dfrac{0.5}{\sigma} = 3.09$

$\sigma = 0.1618$

(*c*) If $P(X < 6.5) = 0.999$, then $P\left(Z < \dfrac{6.5-\mu}{0.25}\right) = 0.999$

Therefore, $\dfrac{6.5-\mu}{0.25} = 3.09$ and $\mu = 5.7275$

Example E3.62

The weight of a mechanical component is normally distributed with a mean of 22 oz and a standard deviation of 0.5 oz.

(*a*) what is the probability that a component weighs more than 23 oz?

(*b*) what must the standard deviation of weight be in order for the company to state that 99.9% of its components are less than 23 oz?

(*c*) if the standard deviation remains at 0.5 oz, what must the mean weight be in order for the company to state that 99.9% of its components are less than 23 oz?

SOLUTION:

Refer to the table in Appendix-*E*.

(*a*) $P(X > 23) = P\left(Z > \dfrac{23-22}{0.5}\right) = P(Z > 2) = 1 - P(Z \leq 2) = 1 - 0.977250 = 0.022750.$

(*b*) If $P(X < 23) = 0.999$, then $P\left[Z < \dfrac{23-22}{\sigma}\right] = 0.999$

Hence, $\dfrac{1}{\sigma} = 3.09$ and $\sigma = \dfrac{1}{3.09} = 0.324$

(*c*) If $P(X < 23) = 0.999$, then $P\left[Z < \dfrac{23-\mu}{0.5}\right] = 0.999$

Hence, $\dfrac{23-\mu}{0.5} = 3.09$

or $\mu = 21.455.$

Example E3.63

Refer to the Example E3.56. Assume that the process has been recalibrated so that mean diameter is now 3.5 cm. To what value must the standard deviation be lowered so that 95% of the diameters will meet the specifications?

SOLUTION:

The specification interval is 3.49 — 3.51 cm. We must find a value for σ so that this interval spans the middle 95% of the population of ball bearing diameters, as shown in Fig. E3.63. The z-score that has 2.5% of the area to its left is $z = -1.96$. The z-score that has 2.5% of the area to its right is $z = 1.96$ (from the symmetry of the curve). It follows that the lower specification limit, 3.49, has a z-score of -1.96, while the upper limit of 3.51 has a score of 1.96. Either of these facts may be used to find σ.

Now $$z = \frac{x-\mu}{\sigma}$$

or $$1.96 = \frac{3.51-3.50}{\sigma}$$

or $$\sigma = 0.005102 \text{ cm}$$

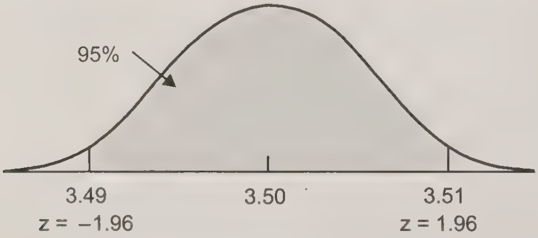


Fig. E3.63

Example E3.64

An experiment needs a 2.41 cm diameter steel rod. Suppose that the diameter of a steel rod has a normal distribution with a mean of 2.41 cm and a standard deviation of 0.01 cm

- *(a)* determine the probability that a diameter is greater than 2.42 cm
- *(b)* what diameter is exceeded by 95% of the samples?
- *(c)* if the specifications require that the diameter is between 2.39 cm and 2.43 cm, what proportion of the samples meet specifications?

SOLUTION:

Refer to the table in Appendix-*E*.

(a) Let X denotes the diameter of the steel rod.

$X \sim N(2.41, 0.01^2)$

$$P(X > 2.42) = 1 - P(X \le 2.42) = 1 - \Phi\left[\frac{2.42-2.41}{0.01}\right] = 1 - \Phi(1) = 0.1587$$

(*b*) $P(X > x) = P\left[Z > \frac{x-2.41}{0.01}\right] = 0.95$

Hence, $\frac{x-2.41}{0.01} = 1.645$ and $x = 2.3936$ cm

(*c*) $P(2.39 < X < 2.43) = \Phi\left[\frac{2.43-2.41}{0.01}\right] - \Phi\left[\frac{2.39-2.41}{0.01}\right] = \Phi(2) - \Phi(-2) = 0.9545$

3.5 APPROXIMATING PROBABILITY DISTRIBUTIONS

In some engineering quality control applications, it is quite useful to approximate one probability distribution by another. Figure 3.8 shows the approximation guide, that is, the conditions under which one distribution may be approximated by another.

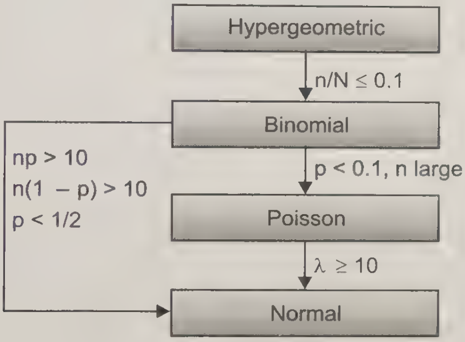


Fig. 3.8: Approximation guide

3.5.1 Binomial Approximation to the Hypergeometric

When the ratio n/N is small, that is, say $n/N \leq 0.1$, then the binomial distribution parameter $p = D/N$ can be used as a good approximation to the hypergeometric distribution. The smaller the function n/N, the better the approximation.

3.5.2 Poisson Approximation to the Binomial

The Poisson distribution can be used as an approximation to the binomial distribution when the sample size n is large and the probability of success p is small (the same applies when q is small, p and q being of course interchangeable), i.e., when the binomial distribution is highly skewed. As a guide, we can say that a good approximation is obtained when $n \geq 20$ and $p \leq 0.05$, and the approximation improves with a decrease in p.

The advantage of using the Poisson distribution to approximate the binomial distribution is that the Poisson distribution, having only one parameter, is well tabulated.

When the parameter is small, say, $p < 0.1$, and n is large, with $\lambda = np$ constant, the Poisson distribution is used as an approximation to the binomial distribution. The larger the value of n and the smaller the value of p, the better the approximation.

3.5.3 Normal Approximation to the Binomial

Recalling that if $X \sim$ Bin (n, p), then $X = Y_1 + Y_2 + \cdots + Y_n$, where $Y_1 + Y_2 + \cdots + Y_n$ is a sample from a Bernoulli (p) population. Hence, X is the sum of the sample observations.

The sample proportion is

$$\hat{p} = \frac{X}{n} = \frac{Y_1 + Y_2 + \cdots + Y_n}{n}$$

which is also the sample mean $\bar{Y}$. The Bernoulli population has mean $\mu = p$ and variance $\sigma^2 = p(1-p)$. It follows from the central limit theorem (see section 3.8) that if the number of trials n is large, then $X \sim N(np, np(1-p))$, and $\hat{p} \sim N(p, p(1-p/n))$.

In the binomial case, the accuracy of the normal approximation depends on the mean number of successes np and the mean number of failures $n(1-p)$. The larger the value of np and $n(1-p)$, the better the approximation. Usually, the normal distribution is used as an approximation to the binomial distribution when np and $n(1-p)$ are both greater than 5. A better and more conservative rule is to use the normal approximation whenever $np > 10$ and $n(1-p) > 10$.

Hence if $X \sim \text{Bin}(n, p)$ and if $np > 10$ and $n(1-p) > 10$, then

$$X \sim N(np, np(1-p)) \qquad \text{approximately}$$

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \qquad \text{approximately}$$

Hence, if X is a binomial random variable with parameter n and p

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

is approximately a standard normal random variable.

If n is large, then one can justify the approximation of the binomial distribution by the normal distribution with mean np and variance $np(1-p)$. Noting that the binomial distribution is discrete and the normal distribution is continuous, we can make a continuity correction such that

$$p(X = a) = \Phi\left[\frac{(a+1/2)-np}{\sqrt{np(1-p)}}\right] - \Phi\left[\frac{(a-1/2)-np}{\sqrt{np(1-p)}}\right] \tag{3.82}$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution. We also write probability statements as

$$P(a \le X \le b) = \Phi\left[\frac{(b+1/2)-np}{\sqrt{np(1-p)}}\right] - \Phi\left[\frac{(a-1/2)-np}{\sqrt{np(1-p)}}\right] \tag{3.83}$$

Although p should be about 0.5, the approximation can be used without excessive loss of accuracy for $0.1 \le p \le 0.9$. If p is close to 0.5 and $n > 10$, the approximation is fairly good. For other values of p, the value of n should be larger.

In general, the approximation is good as long as $np > 5$ for $p \le 0.5$ or when $nq > 5$ when $p > 0.5$.

The normal distribution is also used to approximate the sample fraction nonconforming, $\hat{p}$. The random variable $\hat{p}$ is normally distributed with mean p and variance $p(1 - p)/n$, so that we have

$$p\left[\frac{a}{n} \le \frac{X}{n} \le \frac{b}{n}\right] = P\left[\frac{a}{n} \le \hat{p} \le \frac{b}{n}\right]$$

which indicates that we divide all terms in Eq.(3.83) by n to obtain

$$P\left(\frac{a}{n} \le \hat{p} \le \frac{b}{n}\right) = \Phi\left[\frac{\frac{b+1/2}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}\right] - \Phi\left[\frac{\frac{a-1/2}{n} - p}{\sqrt{\frac{p(1-p)}{n}}}\right] \qquad (3.84)$$

Continuity Correction: The binomial distribution is discrete, while the normal distribution is continuous. The *continuous correction* is an adjustment, made when approximating a discrete distribution with a continuous one that can improve the accuracy of the approximation.

3.5.4 Normal Approximation to the Poisson

If X is a Poisson random variable with $E(X) = \lambda$ and $V(X) = \lambda$

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

is approximately a standard normal random variable. The approximation is good for $\lambda > 10$.

The normal distribution with mean $\mu = \sigma^2 = \lambda$ can be used as an approximation to the Poisson distribution if the mean λ of the distribution is large, say 10 or more.

The continuity correction used is given by

$$P(a \le X \le b) = \Phi\left[\frac{(b+1/2)-\lambda}{\sqrt{\lambda}}\right] - \Phi\left[\frac{(a-1/2)-\lambda}{\sqrt{\lambda}}\right] \qquad (3.85)$$

Example E3.65

The number of widgets made per shift is equal to 1000. It is known that 8% of them are nonconforming. A sample of 20 widgets is taken. What is the probability that 2 or fewer widgets in the sample are nonconforming?

(a) set up the equation using the hypergoemetric distribution

(b) solve it using the binomial approximation

(c) solve it using the Poisson approximation.

SOLUTION:

$$D = 80, N = 1000, n = 20, p = 0.08$$

(a) $$P(X \le 2) = \sum_{x=0}^{2} \frac{\binom{80}{x}\binom{920}{20-x}}{\binom{1000}{20}}$$

(*b*) $P(X \le 2) = \sum_{x=0}^{2} \binom{20}{x} (0.08)^x (0.92)^{20-x}$

$$= \binom{20}{0}(0.08)^0(0.92)^{20} + \binom{20}{1}(0.08)^1(0.92)^{19} + \binom{20}{2}(0.08)^2(0.92)^{18}$$

$$= 0.189 + 0.328 + 0.271 = 0.788$$

(*c*) $\lambda = np = 20(0.08) = 1.6$

$P(x \le 2) = 0.783$

Example E3.66

If 20% of the electronic components made in a company are nonconforming, find the probability that in a random sample of 100 such electronic components selected and using normal approximation.

(*a*) at most 16 will be nonconforming

(*b*) exactly 16 will be nonconforming

(*c*) between 16 and 22 will be nonconforming.

SOLUTION:

Normal approximation

$$np = 100(0.20) = 20 \cdot \sqrt{np(1-p)} = 4$$

(*a*) $P(x \le 16) = \Phi\left[\frac{16.5-20}{4}\right] = \Phi(-0.809) = 1 - 0.809 = 0.191$

(*b*) $P(x = 16) = \Phi\left[\frac{16.5-20}{4}\right] - \Phi\left[\frac{15.5-20}{4}\right] = 0.191 - \phi(-1.125) = 0.191 - (1 - 0.855) = 0.046$

(*c*) $P(16 \le x \le 22) = \Phi\left[\frac{22.5-20}{4}\right] - \Phi\left[\frac{15.5-20}{4}\right] = \phi(0.65) - 0.14 = 0.734 - 0.145 = 0.589$

Example E3.67

Use the normal approximation to the binomial distribution to determine the probability of getting 7 heads and 7 tails in the 14 tosses of a balanced coin.

SOLUTION:

We must thus determine the area under the curve between 6.5 and 7.5. $n = 14$ and $x = 7$, $\mu = 14\left(\frac{1}{2}\right) = 7$ and

$$\sigma = \sqrt{14\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} = \sqrt{3.5} = 1.871$$

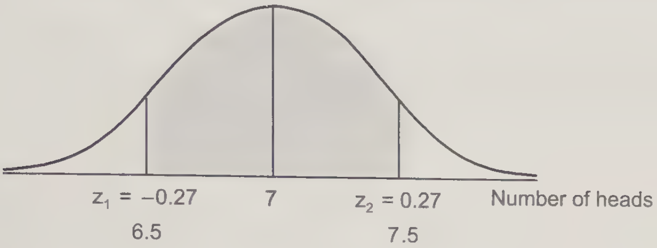


Fig. E3.67

$$z_1 = \frac{6.5-7}{1.871} = -0.27 \text{ and } z_2 = \frac{7.5-7}{1.871} = 0.27$$

The table for normal distribution in Appendix-*E* corresponding to $z = -0.27$ and 0.27 are 0.393580 and 0.60642 and we find that the normal approximation to the probability of "7 heads and 7 tails" is 0.60642 – 0.393580 = 0.2128.

The table for binomial approximation in Appendix-*C* gives 0.2095. Hence, the error of the approximation is

$0.2095 - 0.2128 = -0.0033$ and the percentage error is $\frac{0.0033}{0.2095}(100) = 1.5752\%$ in the absolute value.

Example E3.68

Fifteen per cent of the parts produced in a manufacturing process fail a standardized quality test.

(*a*) using the Poisson formula, find the probability that in a random sample of 100 parts which went through this test exactly 20 will fail.

(*b*) using the Poisson probabilities table, find the probability that the number of parts which fail this test in a randomly selected 100 parts is

- (*i*) at most 9
- (*ii*) 10 to 16
- (*iii*) at least 20.

SOLUTION:

Let x denotes the number of parts in a random sample of 100 which fail the test. Since, on average, 15% of the parts fail the test, $\lambda = 0.15 \times 100 = 15$.

(*a*) $P(\text{exactly 20 fail}) = P(x = 20) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{(15)^{20} e^{-15}}{20!} = 0.0418$

(*b*) (*i*) $P(\text{at most 9 fail}) = P(x \le 9) = P(0) + P(1) + P(2) + P(3) + P(4) + P(5) + P(6) + P(7) + P(8) + P(9)$

$= 0.0 + 0.0 + 0.0 + 0.0002 + 0.0006 + 0.0019 + 0.0048 + 0.0104 = 0.0194 + 0.0324 = 0.0697$

(*ii*) $P(\text{10 to 16 fail}) = P(10 \le x \le 16) = P(10) + P(11) + P(12) + P(13) + P(14) + P(15) + P(16)$

$= 0.0486 + 0.0663 + 0.0829 + 0.0956 + 0.1024 + 0.01024 + 0.0960 = 0.5942$

(*iii*) $P(\text{at least 20}) = P(x \ge 20) = P(20) + P(21) + P(22) + \cdots$

$= 0.0418 + 0.299 + 0.0204 + 0.0133 + 0.0083 + 0.0050 + 0.0029 + 0.0016 + 0.0009$

$+ 0.0004 + 0.0002 + 0.0001 + 0.0001 = 0.1249$

Example E3.69

If the probability is 0.10 that a certain mortgage company will refuse a loan application, use the normal approximation to determine the probability that the mortgage company will refuse at most 40 of 450 mortgage loan applications.

SOLUTION:

$$n = 450, p = 0.10, \mu = np = 450(0.1) = 45,\ \sigma = \sqrt{np(1-p)} = \sqrt{450(0.1)(0.9)} = \sqrt{40.5} = 6.3640$$

$$z = \frac{40.5 - 45}{6.3640} = -0.7071$$

Referring to the table for normal distribution in Appendix-*E* corresponding to $z = -0.7071$, we find that the normal approximation to the probability is 0.22065.

Hence, the probability that the mortgage company will refuse the load application is 0.22065.

Example E3.70

The number of nonconforming manufactured parts per shift is Poisson distributed, with a mean of 16. Find the probability that there are between 14 and 20 nonconforming parts on a shift that is 14, 15, 16, 17, 18, 19 or 20.

(a) use the Poisson distribution

(b) use the normal approximation.

SOLUTION:

(a) $\lambda = 15$

$P[14 \le x \le 20] = F(20) - F(13) = 0.917 - 0.363 = 0.554$

(b) $P[14 \le x \le 20] = \phi\left[\frac{20.5-15}{4}\right] - \phi\left[\frac{13.5-15}{4}\right] = \phi(1.375) - \phi(-0.375) = 0.9154 - 0.3538 = 0.5616$

Example E3.71

Suppose we want to use the normal approximation to the binomial distribution to determine $B(1; 150, 0.05)$

(a) is it justified in using the approximation?

(b) make the approximation.

SOLUTION:

(a) Here $n = 100, p = 0.05$ and mean $= 150(0.05) = 75$ and $n(1-p) = 150 - 7.5 = 142.5$

Yes. It is justified in using the approximation since np is > 5 and $n(1-p)$ is > 100.

(b) $\mu = 7.5, \sigma^2 = np(1-p) = 150(0.05)(0.95) = 7.125$ or $\sigma = 2.67$.

$$z_1 = \frac{0.5 - 7.5}{2.67} = -2.62 \text{ and } z_2 = \frac{1.5 - 7.5}{2.67} = -2.25$$

The entries in table for normal distribution in Appendix-*E* corresponding to $z_1 = -2.62$ and $z_2 = -2.25$, we find that the normal approximation to the probability is

$$0.012224 - 0.004396 = 0.007828$$

Example E3.72

Refer to Example E3.71. Make the Poisson approximation.

SOLUTION:

Here $\lambda = 7.5$ and $P(7.5) = \dfrac{e^{-\lambda}\lambda^x}{x!} = \dfrac{e^{-7.5}(7.5)^1}{1!} = (0.00055)7.5 = 0.0041$

Example E3.73

In a survey it was found that 1.5% of the adult population in a large city suffer from drug addition. Of 100 randomly selected adult population, let X denotes the number who have drug addiction problem.

- *(a)* what are the parameters for the appropriate normal distribution?
- *(b)* what are the parameters for approximating Poisson distribution?
- *(c)* compute the individual probabilities for the binomial distribution in part *(a)*
- *(d)* compute the individual probabilities for the Poisson distribution in part *(b)*. Find the probabilities until they are zero to 4 decimal places.
- *(e)* compare the probabilities found in parts *(c)* and *(d)*.
- *(f)* apply both the binomial probabilities and Poisson probabilities found in parts *(c)* and *(d)* to determine the probability that the number who suffer from drug addiction is exactly 3; between 2 and 5 (inclusive); less than 4% of those surveyed; more than 2. Compare the two results in each case.

SOLUTION:

- *(a)* Here $n = 100$ and $p = 1.5/100 = 0.015$
- *(b)* $\lambda = np = 100(0.015) = 1.5$
- *(c)* The binomial probability function is

$$P(X = x) = \binom{100}{x}(0.015)^x(0.985)100^{-x} \qquad \text{for } x = 0, 1, 2, \ldots, 100$$

Starting at $x = 0$ and evaluating the function until it is zero to four decimal places, we obtain the following Table E3.73(a).

Table E3.73(a)

X	P(X = x)
0	0.2206
1	0.3360
2	0.2532
3	0.1260
4	0.0465
5	0.0136
6	0.0033
7	0.0007
8	0.0001
9	0.0000

(*d*) The Poisson probability function is

$$P(X = x) = e^{1.5}\frac{1.5^x}{x!} \qquad \text{for } x = 0, 1, 2, \dots$$

Starting at $x = 0$ and evaluating this function until it is zero to four decimal places, we obtain the following Table E3.73(b).

Table E3.73(b)

X	P(X = x)
0	0.2231
1	0.3347
2	0.2510
3	0.1255
4	0.0471
5	0.0141
6	0.0035
7	0.0008
8	0.0001
9	0.0000

(*e*) All the probabilities for $x = 0$ to 9 agree to 2 decimal places and are usually at most a couple of thousandths apart. The agreement is really good.

(*f*) See Table E3.73(c).

Table E3.73(c)

Event	Binomial probability	Poisson probability
$X = 3$	0.1260	0.1255
$2 \le X \le 5$	0.4393	0.4377
$X < 4$	0.9358	0.9343
$X > 2$	0.1902	0.1912

Each pair of probabilities in the above Table E3.73(c) agree to within 0.0016 or less.

3.6 CHEBYSHEV'S THEOREM

The standard deviation σ of a random variable X measures the weighted spread of the values of X about the mean μ. For smaller σ, we would expect that X will be closer to μ. A more precise statement of this expectation is given by Chebyshev (1821–1894) which is stated as follows:

Let X be a random variable with mean μ and standard deviation σ. Then, for any integer k, the probability that a value of X lies in the interval $[\mu - k\sigma, \mu + k\sigma]$ is at least $1 - \dfrac{1}{k^2}$.

Hence, $$P[\mu - k\sigma \le X \le \mu + k\sigma] \ge 1 - \frac{1}{k^2} \qquad (3.86)$$

Proof: By definition

$$\sigma^2 = \text{Var}(X) = \Sigma(x_i - \mu)^2 p_i \qquad (3.87)$$

Now, deleting all terms for which $|x_i - \mu| \leq k\sigma$ and denoting the summation of the remaining terms by $\Sigma^*(x_i - \mu)^2 p_i$. then, we have

$$\sigma^2 \geq \Sigma^*(x_i - \mu)^2 p_i \geq \Sigma^* k^2\sigma^2 p_i = k^2\sigma^2 \Sigma^* p_i = k^2\sigma^2 p(|X - \mu| > k\sigma) \tag{3.88}$$

or

$$= k^2\sigma^2[1 - p(|X - \mu| \leq k\sigma)] = k^2\sigma^2[1 - P(\mu - k\sigma \leq X \leq \mu + k\sigma)] \tag{3.89}$$

when $\sigma > 0$, and dividing Eq. (3.89) by $k^2\sigma^2$, we get

$$\frac{1}{k^2} \geq 1 - P(\mu - k\sigma \leq X \leq \mu + k\sigma) \tag{3.90}$$

or

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \leq 1 - \frac{1}{k^2} \tag{3.91}$$

Equation (3.91) is the Chebyshev's inequality for $\sigma > 0$.

If $\sigma = 0$, then $x_i = \mu$ for all $p_i > 0$, and we have

$$P[\mu - k(0) \leq X \leq \mu + k(0)] = P(X = \mu) = 1 > 1 - \frac{1}{k^2} \tag{3.91a}$$

Chebyshev's theorem gives a lower bound for the area under a curve between two points that are an opposite sides of the mean and at the same distance from the mean. Chebyshev's theorem is stated as follows: For any number k greater than 1, at least $\left(1 - \frac{1}{k^2}\right)$ of the data values lie within k standard deviations of the mean. Chebyshev's theorem is illustrated in Fig. 3.9.

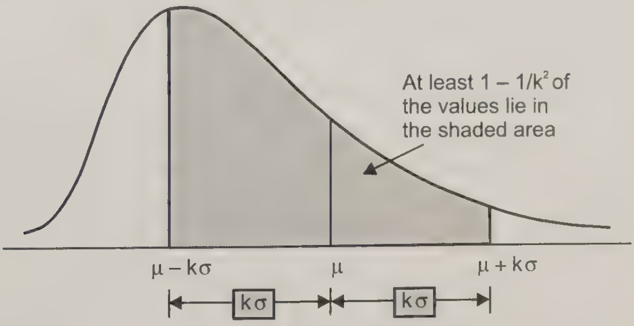


Fig. 3.9: Chebyshev's theorem

For $k = 2$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 0.75 \text{ or } 75\%$$

That is, at least 75% of the values of a data set lie within two standard deviations of the mean as shown in Fig. 3.10.

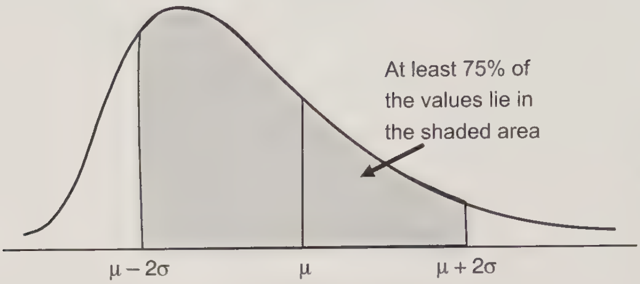


Fig. 3.10: Percentage of values within two standard deviations of the mean

For $k = 3$

$$1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 0.89 \text{ or } 89\%$$

When $k = 3$, this inequality shows that regardless of the assumed distribution of X, the probability is no more than $1/3^2 = 1/9$ that the random variable takes on a value more than three standard deviations away from its mean (see Fig. 3.11).

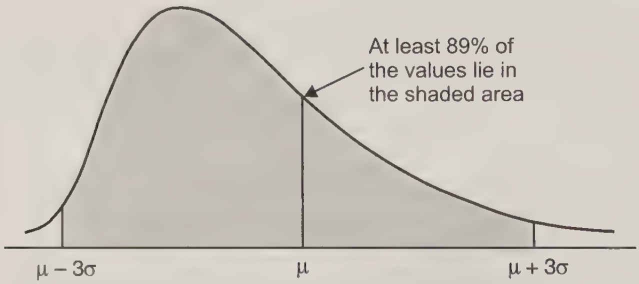


Fig. 3.11: Percentage of values within three standard deviations of the mean

Chebyshev's inequality is applicable to both sample and population data. Chebyshev's inequality is also applicable to distribution of any shape. It should be noted here that Chebyshev's inequality can be used only for $k > 1$ (since when $k = 1$, $1 - \frac{1}{k^2} = 0$ and when $k < 1$, $1 - \frac{1}{k^2}$ is negative).

Example E3.74

The mean and standard deviation for the final examination scores in *Economics* course are 80 and 7.5 respectively. Determine the percentage of students who scored between 65 and 95, using Chebyshev's theorem.

SOLUTION:

From the given data

Mean $\mu = 80$ and standard deviation $\sigma = 10$.

Each of the two points, 60 and 95, is 15 units away from the mean.

Hence,
$$k = \frac{15}{\sigma} = \frac{15}{7.5} = 2$$

$$1-\frac{1}{k^2}=1-\frac{1}{2^2}=0.75 \text{ or } 75\%$$

That is, at least 75% of the students scored between 60 and 95 as shown in Fig. E3.74.

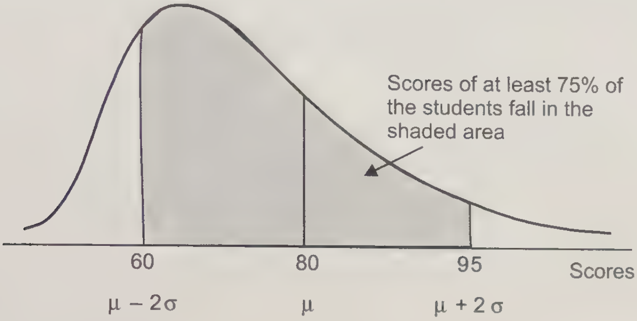


Fig. E3.74: Percentage of students who scores between 60 and 95

Example E3.75

A random variable X has mean $\mu = 35$ and standard deviation $\sigma = 2$. Apply Chebyshev's inequality to estimate

(*a*) $P(X \leq 45)$

(*b*) $P(X \geq 30)$

Solution:

Chebyshev's inequality states:

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) \geq 1-\frac{1}{k^2}$$

(*a*) Substituting $\mu = 35$ and $\sigma = 2$ in $\mu + k\sigma$, we have

$$35 + 2k = 45 \text{ or } k = 5$$

Hence, $$1-\frac{1}{k^2}=1-\frac{1}{5^2}=0.96$$

Similarly, $\mu - k\sigma = 35 - 10 = 25$

Chebyshev's inequality gives

$P(25 \leq X \leq 45) \geq 0.96$.

The event corresponding to $X \leq 45$ contains as a subset the event corresponding to $25 \leq X \leq 45$. Hence

$$P(X \leq 45) \geq P(25 \leq X \leq 45) \geq 0.96$$

Therefore, the probability that X is less than or equal to 45 is at least 96%

(*b*) Substituting $\mu = 35$ and $\sigma = 2$ in $\mu + k\sigma$, we have

$$35 - 2k = 30 \text{ for } k \text{ or } k = 2.5$$

$$1-\frac{1}{k^2}=1-\frac{1}{2.5^2}=0.84$$

Since $\mu + k\sigma = 35 + 5 = 40$, Chebyshev's inequality gives

$$P(30 \le X \le 40) \ge 0.84$$

The event corresponding to $X \ge 30$ contains as a subset the event corresponding to $30 \le X \le 40$. Hence

$$P(X \ge 45) \ge P(30 \le X \le 40) \ge 0.84$$

Thus, the probability that X is greater than or equal to 30 is at least 84%.

Example E3.76

A random variable X has mean $\mu = 85$ and standard deviation $\sigma = 5$.

(a) what inferences can be made from Chebyshev's inequality for $k = 2$ and $k = 3$?

(b) estimate the probability that X lies between 65 and 105.

(c) determine an internal $[a, b]$ about the mean for which the probability that X lies in the interval is at least 99%.

SOLUTION:

(a) We set $k = 2$ and obtain $\mu - k\sigma = 85 - 2(5) = 75$ and $\mu + k\sigma = 85 + 2(5) = 95$.

We can therefore conclude from Chebyshev's inequality that the probability that a value of X lies between 75 and 95 is at least $1 - \frac{1}{2^2} = 0.75$.

$$P(75 \le X \le 95) \ge 0.75$$

If $k = 3$, we find that the probability that X lies between 70 and 100 is at least $1 - \frac{1}{3^2} = 8/9 = 0.889$

(b) Here $k\sigma = 20$ since $85 - 20 = 65$ and $85 + 20 = 105$ or $k(5) = 20$ or $k = 4$.

Hence, by Chebyshev's inequality

$$P(65 \le X \le 105) \ge 1 - \frac{1}{4^2} = 0.94$$

Thus, the probability X lies between 65 and 105 is at least 94%.

(c) Setting $1 - \frac{1}{k^2} = 0.99$, we get $1 - 0.99 = 1/k^2$ or $k = 10$.

Hence the interval is $[85 - 10(5), 85 + 10(5)] = [35, 135]$.

3.7 EMPIRICAL RULE

If the distribution of the data is approximately bell-shaped, then empirical rule applies. The empirical rule is stated as follows:

For a bell-shaped distribution, approximately

1. 68% of the observations lie within one standard deviation of the mean.
2. 95% of the observations lie within two standard deviations of the mean.
3. 99.75% of the observations lie within three standard deviations of the mean.

The empirical rule is illustrated in Fig. 3.12. The bell-shaped distribution, also known as the normal distribution has been described earlier. The empirical rule is applicable to population data as well as sample data.

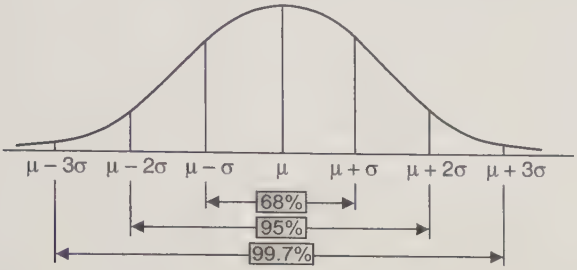


Fig. 3.12: Empirical rule

Example E3.77

The weight distribution of a sample of 10,000 male students in a particular university is bell-shaped with a mean of 40 kg and a standard deviation of 5 kg. Find the percentage of students who weigh approximately 30 to 50 kg.

SOLUTION:

From the given data, we have for the sample

Mean $\bar{x}$ = 40 kg and standard deviation $s = 5$ kg.

Each of the two points, 30 and 50, is 10 units away from the sample mean of 40 kg. Dividing 10 by the standard deviation (5 kg), we have $\frac{10}{\sigma} = \frac{10}{5} = 2$. The distance between 30 and 40 and between 40 and 50 kg is each equal to $2s$ as shown in Fig. E3.77, the area from 30 to 50 kg is the area from $\bar{x} - 2s$ to $\bar{x} + 2s$. Applying the empirical rule, approximately 95% of the male students in the sample weigh between 30 and 50 kg.

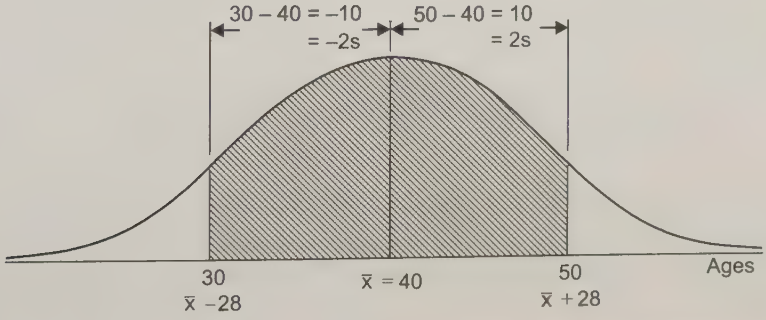


Fig. E3.77: Percentage of male students who weigh between 30 and 50 kg

3.8 THE CENTRAL LIMIT THEOREM

Many commonly used statistical methods depend on the central limit theorem. The central limit theorem is very important result in statistics. The central limit theorem states that if a large enough sample is drawn from a population, then the distribution of the sample mean is approximately normal, no matter what population the sample was drawn from. If the variable is normally distributed, so is the variable mean, $\bar{x}$. This key fact also holds approximately if x is not normally distributed, provided only that the sample size is relatively large.

Let $X_1, X_2, \ldots, X_n$ be a simple random sample from a population with mean μ and variance σ.

Let $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$ be the sample mean.

Let $S_n = X_1 + X_2 + \cdots + X_n$ be the sum of the sample observations.

Then, if n is sufficiently large,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \text{ approximately}$$

and $$S_n \sim N(n\mu, n\sigma^2) \text{ approximately}$$

The statement of the *central limit theorem* specifies that $\mu_{\bar{x}}$ and $\sigma^2_{\bar{x}} = \dfrac{\sigma^2}{n}$, which hold for any sample mean. The sum of the sample items is equal to the mean multiplied by the sample size, that is, $S_n = n\bar{x}$. It follows that $\mu_{S_n} = n\mu$ and $\sigma^2_{S_n} = n^2\sigma^2/n = n\sigma^2$.

Hence, for a relatively large sample size, the variable is approximately normally distributed, regardless of the distribution of the variable under consideration. The approximation becomes better with increasing sample size.

The further the variable under consideration is from being normally distributed, the larger the sample size must be for a normal distribution to provide an adequate approximation to the distribution of $\bar{x}$. Generally, however, a sample size of 30 or more ($n \geq 30$) is large enough.

The proof of the central limit theorem is difficult and very complex to be included here.

Summarising, the central limit theorem says that if $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2, and if $\bar{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$, is the standard normal distribution. The normal approximation for $\bar{X}$ depends on the sample size n. In many cases of practical interest, if $n \geq 30$, the normal approximation will be satisfactory regardless of the shape of the population. If $n < 30$, the central limit theorem will work if the distribution of the population is not severely non normal.

Example E3.77

A mechanical component has tensile strength that is normally distributed with mean 75.5 MPa and standard deviation 3.5 MPa. Find the probability for a random sample of $n = 6$, these component specimens will have sample mean tensile strength that exceeds 75.75 MPa.

SOLUTION:

Given $\mu_{\bar{X}} = 75.5$ MPa

Therefore, $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{3.5}{\sqrt{6}} = 1.429$

$$P(\bar{X} \geq 75.75) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right) = P(Z \geq 0.175) = 1 - P(Z \leq 0.175)$$

$$= 1 - 0.56945 = 0.43055$$

Example E3.78

A normal population has mean 80 and variance 25. How longer must the random sample be if we want the standard error of the sample average to be 1.5?

SOLUTION:

Given $\sigma^2 = 25$ or $\sigma = 5$ and $\sigma_{\bar{X}} = 1.5$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{5}{\sqrt{n}} = 1.5$$

or $$n = \left(\frac{\sigma}{\sigma_{\bar{X}}}\right)^2 = \left(\frac{5}{1.5}\right)^2 = 11.11 \text{ or } \approx 12$$

PROBLEMS

Section 3.1 Random Variables

P3.1 The frequency distribution of the number of orders received per day by a distribution company during the past 3 months (90 days)

Number of orders received per day	2	3	4	5	6
Number of days	11	20	23	22	14

(a) construct a probability distribution table for the number of orders received per day

(b) are the probabilities in *(a)* exact or approximate?

(c) if x denotes the number of orders received on any given day, calculate the following probabilities

(i) $P(x = 3)$
(ii) $P(x \geq 3)$
(iii) $P(2 \leq x \leq 4)$
(iv) $P(x < 4)$.

P3.2 The following table gives the probability distribution of a discrete random variable x.

x	0	1	2	3	4	5	6
P(x)	0.11	0.14	0.18	0.15	0.12	0.09	0.05

Find the following probabilities:

(a) $P(x = 4)$

(b) $P(x \geq 3)$

(c) $P(x < 3)$

(d) $P(3 \leq x \leq 5)$

P3.3 The following table gives the probability distribution of automobiles sold on a given day in an automobile dealership.

Automobile sold	0	1	2	3	4	5	6
Probability	0.05	0.10	0.20	0.30	0.20	0.15	0.05

Calculate the mean and standard deviation for this probability distribution.

P3.4 Table P3.4 gives the probability distribution where X represents the number of times the manufacturing process is recalibrated during a week whenever the quality of the component produced falls below certain specifications. Assume that X has the probability mass function as given in Table P3.4.

Table P3.4

x	0	1	2	3	4
p(x)	0.35	0.25	0.20	0.15	0.05

(a) determine the mean of X

(b) determine the variance and standard deviation for the random variable X.

P3.5 Table P3.5 shows the probability distribution of the number of breakdowns per week for a machine.

Table P3.5

Breakdowns per week	0	1	2	3
Probability	0.15	0.20	0.35	0.30

(a) represent the probability graphically

(b) find the probability that the number of breakdowns for this machine during a given week is

(i) exactly 2

(ii) 0 to 2

(iii) more than 1

(iv) at most 1

P3.6 For each of the following, indicate whether it is or is not a random variable.

- (*a*) the number of automobiles sold at a dealership during a given month
- (*b*) the height of a person
- (*c*) the number of heart beats per minute
- (*d*) the price of an automobile
- (*e*) the time needed to complete a written examination
- (*f*) the number of automobiles a household owns
- (*g*) the number of tails obtained in four tosses of a balanced coin.

P3.7 The number of telephone calls received in a particular sales office between 10:00 am and 11:00 am has the probability function given in Table P3.7.

Table P3.7

Number of telephone calls received, x	0	1	2	3	4	5	6
Probability, P(x)	0.05	0.20	0.25	0.20	0.10	0.15	0.05

- (*a*) is this a probability function?
- (*b*) find the probability that there will be three or more calls
- (*c*) find the probability that there will be an even number of calls.

P3.8 Table P3.8 gives the probability distribution of x, where x denotes the number of defective mechanical components produced in a manufacturing company.

Table P3.8

x	0	1	2	3	4	5
P(x)	0.02	0.20	0.30	0.30	0.10	0.08

Determine the mean and standard deviation of x.

P3.9 Table P3.9 gives the number of telephone calls received per hour in an office and the distribution.

Table P3.9

Number of telephone calls received, x	10	11	12	13	14	15
f (x)	0.08	0.15	0.30	0.20	0.20	0.07

Determine the mean and standard deviation of the number of telephone calls received per hour in that office.

P3.10 Table P3.10 gives the results of the experiment of rolling a die with the discrete random variable number of dots.

Table P3.10

Number of dots, x	1	2	3	4	5	6
Probability, f (x)	1/6	1/6	1/6	1/6	1/6	1/6

Determine the mean and variance of that random variable x.

P3.11 The time to failure of an electronic component is a continuous random variable known to have the density function $0.5e^{-0.5t}$ where t is in years. What is the probability that this component will fail within the first year of operation?

P3.12 Find the mean μ for the probability density function, $\rho(x)$, of the life of a projector bulb, random variable x, is given as:

and
$$\rho(x) = \begin{cases} 0 & ; \quad \text{for } x < 0 \\ 1/900\, e^{-x/900} & ; \quad \text{for } x \geq 0 \end{cases}$$

P3.13 The density function for a continuous random variable x is given as:

$f(x) = 0.25(x - 2)$ for $2 \leq x \leq 5$

Sketch the density and distribution functions.

P3.14 The daily consumption of an electric power of a certain machine (in units of power) is a random variable whose probability density is given by:

$$f(x) = \begin{cases} 1/9\, x e^{-x/3} ; & \text{for } x > 0 \\ 0 ; & \text{for } x \leq 0 \end{cases}$$

Determine the probabilities that on a given day

(a) the consumption of this machine is no more than 6 units

(b) the power supply is inadequate in the daily capacity if the supply is 9 units.

P3.15 The total lifetime (in years) of a certain machine is a random variable whose distribution function is given by:

$$f(x) = \begin{cases} 0 ; & \text{for } x \leq 5 \\ 1 - 25/2^2 ; & \text{for } x > 5 \end{cases}$$

Find the probabilities that such a machine system will have life

(a) beyond 10 years

(b) less than 8 years

(c) anywhere from 12 to 15 years.

P3.16 The probability density function of X is given by

$$f(x) = \begin{cases} 1.25(1 - x^4) ; & 0 < x < 1 \\ 0 ; & \text{otherwise} \end{cases}$$

where the random variable X denotes the clearance (in mm). The clearance is the difference between the radius of the hole drilled in a flat sheet-metal plate and a shaft inserted through the hole. Components with clearances larger than 0.8 mm are to be scrapped. What proportion of components are scrapped?

P3.17 Refer to Problem P3.16. Determine the mean clearance and the variance of the clearance.

P3.18 Determine the mean and variance for the function $f(x)$

$$f(x) = \begin{cases} \dfrac{1}{b - a} ; & a \leq x \leq b \\ 0 ; & \text{otherwise} \end{cases}$$

where X is the uniform random variable on the interval $[a, b]$.

P3.19 A certain radioactive mass emits alpha particles from time to time. The time between emissions (in seconds) is random with the probability density function given by

$$f(x) = \begin{cases} 0.1e^{-0.1x} & ; \ x > 0 \\ 0 & ; \ x \leq 0 \end{cases}$$

Determine the median time between emissions.

P3.20 Determine the mean and variance for the following continuous random variable X with probability density function $f(x)$ given by

$$f(x) = \begin{cases} \frac{1}{\lambda} e^{-x/\lambda} & ; \ x \geq 0 \text{ and } \lambda > 0 \\ 0 & ; \ \text{otherwise} \end{cases}$$

Section 3.2 Permutations and Combinations

P3.21 (*a*) Determine the number of permutations of the 26 letters of the alphabet.

(*b*) Determine the number of ways of 6 distinct products that can be lined up on a display shelf.

P3.22 A committee is to be formed consisting of 2 men and 2 women. There are 6 men and 7 women qualified and available to fill this position. How many different committees can be formed out of the finalists?

P3.23 Develop a table of the number of possibilities when 2, 3, 4 and 5 letters of the alphabet are used. Generate the possibilities. English alphabet letters are used with and without repetition.

P3.24 How many 5 digit numbers can be formed with the 10 digits 0, 1, 2, 3, ..., 9 if

(*a*) repetitions are allowed

(*b*) repetitions are not allowed

(*c*) the last digit must be zero and repetitions are not allowed.

P3.25 How many ways can an executive committee of 5 can be chosen from a pool of 15 members?

P3.26 A company wants to purchase 4 electronic systems. After all the system models were reviewed, 8 foreign made and 10 U.S. made systems were considered to satisfy all the security requirements for the company.

(*a*) if the systems are chosen at random, find the probability that 2 of the systems selected are foreign made.

(*b*) what is the probability that all the four systems selected are U.S. made?

(*c*) what is the probability that all of the 4 systems selected are foreign made?

(*d*) what is the probability that at least 2 of the systems are U.S. made?

Permutations

P3.27 A pianist knows five pieces but will have enough stage time to play only four of them. Pieces played in a different order constitute a different program. How many different programs can be arranged?

P3.28 There are six persons on a sinking boat. There are four life jackets on board. How many combinations of survivors are there?

P3.29 A violinist knows four pieces but will have enough time to play only three of them. Pieces played in a different order constitute a different program. Determine the number of different programs that can be arranged.

P3.30 A bin contains 5 white balls, 2 red balls, and 3 green balls. Find the probability of getting either a white ball or a red ball in one drawn from the bin.

P3.31 Five defective parts were unintentionally mixed with 45 non-defective (good) ones. After a thorough mixing, 5 parts are picked simultaneously from the collection of 50.

(a) determine the probability that all 5 parts selected are non-defective ones

(b) what is the probability that no more than 1 part selected is defective?

(c) what is the probability that all 5 parts selected are defective?

P3.32 From 12 items, in how many ways can a selection of 6 be made?

(a) when a specified item is always included

(b) when a specified item is always excluded.

P3.33 Out of 6 manufacturing engineers and 9 design engineers, a committee consisting of 3 manufacturing engineers and 5 design engineers is to be formed. In how many ways can this be done if:

(a) any manufacturing engineer and any design engineer can be included

(b) one particular design engineer must be on the committee

(c) two particular manufacturing engineers cannot be on the committee.

P3.34 Determine the number of ways of splitting 10 persons into 2 groups containing 4 and 6 persons respectively.

P3.35 A committee of 2 men and 3 women needs to be formed out of 7 men and 7 women. Determine the number of ways this can be done if

(a) any man and any woman can be included

(b) one particular woman must be on the committee

(c) two particular men cannot be on the committee.

P3.36 Three cards are drawn in succession from a deck of 52 cards. Determine the number of ways this can be done

(a) with replacement

(b) without replacement.

P3.37 Determine the number of committees of 3 that can be formed from 8 persons.

P3.38 Determine the number of combinations of 4 taken 3 at a time.

P3.39 Find the number of ways of selecting two objects out of four objects.

P3.40 If 3 persons are to be selected randomly from 5 persons for a committee, determine the different possible combinations.

Section 3.3 Discrete Distributions

P3.41 A lot contains 140 electronic components and 20 are selected without replacement for quality testing.

(a) if 20 components are defective, what is the probability that at least one defective component is in the sample?

(b) if 5 components are defective, what is the probability that at least one defective component appears in the sample?

P3.42 A lot of mechanical parts contains 24 in a package. It was determined to test 4 parts in each package and if all 4 parts pass the tests, then they are all accepted. If in a package of parts in which 3 are non-conforming, what is the probability of rejection of these parts?

P3.43 A die is rolled three times. Find the probability of getting one 4 in the three rolls. Also find the probability of getting two 4'*s*, three 4'*s*, and no 4'*s* in three rolls.

P3.44 In a production process the defective rate is 15 per cent. Assuming a random sample of 10 items is drawn from this process, find the probability that two of them are defective.

P3.45 Batches of 50 shock absorbers from a production process are tested for conformance to quality requirements. The mean number of non-conforming absorbers in a batch is 5. Assume that the number of non-conforming shock absorbers in a batch, denoted as x, is a binomial random variable.

(a) find n and p

(b) find $p(x \leq 2)$

(c) find $p(x \geq 49)$.

P3.46 A production process manufactures certain mechanical components for a machine system. On average, 1.5% of the components will not perform up to specifications. When a shipment of 100 components is received at the plant, they are tested, and if more than 2 are defective, the shipment is returned to the manufacturer. What is the probability of returning a shipment?

P3.47 A quality test engineer claims that 1 in 10 of certain manufactured parts is due to material defects. Using binomial distribution and rounding to four decimals, find the probability that at least of 3 out of 5 of the tested parts are due to material defects.

P3.48 At the ABC House Delivery Service, providing high-quality service to its customers is the top priority of the company. The company guarantees a refund of all charges if a package it is delivering does not arrive at it's destination by the specified time. It is known from past data that despite all efforts, 3% of the packages mailed through this company do not arrive at their destinations within the specified time. A corporation mailed 10 packages through ABC House Delivery Service on Monday.

(a) find the probability that exactly one of these 10 packages will not arrive at it's destination within the specified time.

(b) find the probability that at most one of these 10 packages will not arrive at it's destination within the specified time.

P3.49 Five per cent of a large batch of high-strength steel components purchased for a mechanical system are defective.

(a) if seven components are randomly selected, find the probability that exactly three will be defective

(b) find the probability that two or more components will be defective.

P3.50 The number of customers arriving a bank in the next period is a Poisson distribution having a mean of eight. Find the probability that exactly six customers will arrive in the next period.

P3.51 If the probability of a concrete beam failing in compression is 0.05, use the Poisson approximation to obtain the probability that from a sample of 50 beams

(a) at least three will fail in compression

(b) no beam will fail in compression.

P3.52 The number of defects on an electronic component which is used in a computerized system has been found to follow the Poisson distribution with $\lambda = 3$. Find the probability that a randomly selected electronic component will have two or less defects.

P3.53 The number of telephone calls made to a certain company's operator is a Poisson random variable with a mean of 5 calls per hour.

(a) what is the probability that 5 calls are received in one hour?

(b) what is the probability that 10 calls are received in 1.5 hour?

(c) what is the probability that less than 2 calls are received in 1 – 1/2 hours?

P3.54 A photocopying machine in an office breaks down an average of three times per month. Using the Poisson probability distribution formula, find the probability that during the next month this machine will have

(a) exactly two breakdowns

(b) at most one breakdown.

P3.55 The proportion of mechanical manufactured parts that are non-conforming is 0.04. Obtain the Poisson approximation to the binomial distribution for the probability of three or fewer non-conforming parts in a sample of 100.

P3.56 An examination consists of five questions, and to pass the examination a student has to answer at least four questions correctly. Each question has three possible answers, of which only one is correct. If a student guesses on each question, what is the probability that the student will pass the test?

P3.57 The probability that a person undergoes a heart operation will recover is 0.6. Determine the probability that of the six patients who undergo similar heart operation:

(a) none will recover

(b) all will recover

(c) half will recover

(d) at least half will recover.

P3.58 In a certain manufacturing process, it was found from quality control inspection that 20% of the machine components produced by the process are defective. Determine the probability that out of 4 machine components selected at random *(a)* 1, *(b)* 0, *(c)* at most 2 machine components will be defective.

P3.59 Given that the probability of an individual patient suffers a bad reaction from injection of a particular serum is 0.001. Determine the probability that out of 2000 individual patients.

(a) exactly 3 individuals will suffer a bad reaction

(b) more than 2 individuals will suffer a bad reaction. Use Poisson distribution.

P3.60 Use binomial distribution and repeat Problem 3.59.

Section 3.4 Continuous Probability Distributions

P3.61 The inside diameter of a finished shaft of uniform diameter is normally distributed with a mean of 4.50 cm and a standard deviation of 0.01 cm. What is the probability of obtaining a diameter exceeding 4.52 cm?

P3.62 The resistance of a foil strain gauge is normally distributed with a mean of 100 ohms and a standard deviation of 0.8 ohm. The specification limits are 100 ±1.0 ohms. What percentage of gauges will be defective?

P3.63 The measurement of the diameter of a special steel pipe is normally distributed with a mean of 5.01 cm and standard deviation of 0.03 cm. The specification limits are 5.00 ±0.05 cm. What percentage of pipes is not acceptable?

P3.64 The yield strength of metal component manufactured is $N(192.36)$. A purchaser of the metal components requires strength of at least 180 psi. The probability that a metal component will meet or exceed the specifications is given by $P(x \geq 180)$. Determine the probability that a metal component meets or exceeds the specifications.

P3.65 Find the probability that the yield strength of the metal components in Problem P3.64 is between 180 and 198 psi.

P3.66 The diameter of a machine shaft produced in a manufacturing company is normally distributed with a mean diameter of 0.001 inches and a standard deviation of 0.0002 inches.

(a) what is the probability that the diameter of the shaft exceeds 0.0013 inches?

(b) what is the probability that a diameter of the shaft is between 0.0007 and 0.0013 inches?

(c) what standard deviation of diameters is needed so that the probability in part (*b*) is 0.995?

P3.67 The mass, m, of a particular machine part is normally distributed with a mean of 66 kg and a standard deviation of 5 kg.

(a) what percentage of the parts will have a mass less than 72 kg?

(b) what percentage of parts will have a mass in excess of 72 kg?

(c) what per cent of the parts will have a mass between 61 and 72 kg?

P3.68 It was determined experimentally that the load X required to break a plate is normally distributed with mean 2.5 and standard deviation 0.24. Determine the probability that

(a) the plate breaks at a load of 2.61 or less

(b) the plate breaks at a load of more than 2.39

(c) the breaking load is at least 2.86

(d) the breaking load is in the interval (2.61, 2.86).

P3.69 Use the Table in Appendix-*E* to determine the following probabilities for the standard normal random variable z.

(a) $P(z < 3.0)$

(b) $P(z > -2.15)$

(c) $P(-3 < z < 3)$

(d) $P(0 < z < 1)$

(e) $P(z > 3)$

P3.70 Assume z has a standard normal distribution. Use Appendix-*E* table to determine the value for z that solves each of the following:

(a) $P(Z < z) = 0.5$

(b) $P(Z > z) = 0.9$

(c) $P(-z < Z < z) = 0.99$

(d) $P(-z < Z < z) = 0.68$

(e) $P(-1.24 < Z < z) = 0.8$

P3.71 The diameter of a component in a machine system is normally distributed with mean 0.2508 cm and standard deviation 0.0005 cm. The specifications on the component are 0.25 ±0.0015 cm. Determine the proportion of components confirms to specifications.

P3.72 The mass, μ, of a particular electronic component is normally distributed with a mean of 66 g and a standard deviation of 5 g. Determine

(a) the per cent of components that will have a mass less than 72 g

(b) the per cent of components that will have a mass in excess of 72 kg

(c) the per cent of components that will have a mass between 61 and 72 g.

P3.73 Customers buying copper rods supplied by a certain manufacturer require that the rods be between 9.9 cm and 10.5 cm, inclusive. The manufacturing process is such that the actual rod lengths are well approximated by a normal distribution with mean 10.1 cm and standard deviation 0.20 cm. Determine the percentage of the manufacturer's production is acceptable to the customer.

P3.74 Refer to Problem P3.73 and determine what rod length is exceeded by 95% of the manufacturer's product.

Section 3.5 Approximating Probability Distributions

P3.75 In a production lot of 100 components, five of them are found to be non-conforming. Approximate the probability that a random sample of 10 components contains no more than 1 non-conforming component (hypergeometric distribution) by the binomial distribution.

P3.76 The proportion of components manufactured that are non-conforming is 0.04. The probability of three of fewer non-conforming components in a random sample of 100 is given by the binomial distribution. Make an approximation using Poisson distribution.

P3.77 Out of a batch of 2800 electronic components received 25% of them were defective or non-conforming. In a sample of 50 randomly selected, the probability that between 12 and 14 are given by the binomial distribution. Use normal distribution to this binomial distribution.

P3.78 A lot of electronic components are known to contain $p = 0.12$ non-conforming. In a sample of 100, the probability that the fraction non-conforming is $0.10 \leq p \leq 0.20$ is given by the normal distribution. Make an approximation.

P3.79 The number of telephone calls received in a specified time by an operator in a manufacturing company is Poisson distribution with mean $\lambda = 14$. Determine the probability that between 10 and 18 calls are received by the operator in that specified time using normal approximation.

P3.80 Determine the probability that in a sample of 10 machine components chosen at random, exactly two will be defective by using

(a) the binomial distribution

(b) the Poisson approximate to the binomial distribution.

Given that 10% of the machine components produced in that manufacturing process are defective.

P3.81 In some manufacturing process that a production of 200 components contains 5 components that do not meet the quality specifications. Determine the probability that a random sample of 10 components will contain no non-conforming components. Use binomial approximation to the hypergeometric.

P3.82 Assume that an examination has 10 questions of the type true or false. If the student taking such an examination guesses at all 10 questions, determine the probability that the student answers either seven or eight questions correctly.

P3.83 *(a)* If a fair coin is tossed 100 times, use the normal curve to approximate the probability that the number of heads is between 35 and 45 inclusive.

(b) If a fair coin is tossed 100 times, use the normal curve to approximate the probability that the number of heads is between 35 and 45 exclusive.

P3.84 In a certain university, 25% of the students are over 21 years of age. In a sample of 400 students, what is the probability that more than 110 of them are over 21 years of age?

P3.85 A soft drink company conducted a taste survey marketing a new soft drink. The results of this survey showed that 70% of the people who tried the drink liked it. Encouraged by this result, the company decided to market the new soft drink. Assume that 70% of all people like this drink. On a certain day, 100 customers bought this soft drink.

(a) find the probability that exactly 65 out of 100 customers will like this drink

(b) find the probability that exactly 60 or less of the 100 customers will like this drink

(c) find the probability that exactly 75 to 80 of the 100 customers will like this drink

P3.86 The length of a bolt manufactured in a certain manufacturing process has a mean of 50 mm and standard deviation of 0.45 mm. Determine the largest possible value for the probability that the length of the bolt is outside the interval 49.1 — 50.9 mm.

P3.87 In a mechanical engineering design class, the mean for the final examination scores is 75 and the standard deviation is 5. Using Chebyshev's theorem, find the percentage of students who scored between 60 and 90.

P3.88 Given that X is a random variable with the $N(\mu, \sigma^2)$ distribution. Denoting the standardised X by Z, make the comparisons between the actual and the Chebyschev's bounds for $Z = 1, 2,$ and $3(k = 1, 2,$ and $3)$.

P3.89 In mechanics class, the mean for the midterm scores is 65 and the standard deviation is 8. Using Chebyshev's theorem, find the percentage of students who scored between 49 and 81.

P3.90 The 2007 gross sales of all firms in a large city have a mean of 3.3 million and a standard deviation of 0.6 million. Using Chebyshev's theorem, find at least what percentage of firms in this city had 2007 gross sales of

(a) \$2.1 to \$4.5 million

(b) \$1.8 to \$4.8 million

(c) \$1.5 to \$5.1 million

P3.91 *(a)* Let X be a random variable with mean $\mu = 50$ and standard deviation $\sigma = 6$. Use Chebyshev's inequality to find a value for b for which $P(50 - b \leq X \leq 50 + b) \geq 0.95$.

(b) Let X be a random variable with mean $\mu = 70$ and unknown standard deviation σ. Use Chebyshev's inequality to find a value for σ for which $P(65 \leq X \leq 75) \geq 0.90$.

P3.92 The age distribution of a sample of 6000 persons is bell-shaped with a mean of 50 years and a standard deviation of 10 years. Determine the approximate percentage of people who are 30 to 70 years old.

P3.93 A manufacturing company manufactures steel rods of 100 cm length and a standard deviation of 10 cm. The distribution of the rod is normal. Find the probability that a random sample of $n = 25$ rods will have an average length less than 95 cm.

P3.94 The following table gives the probability mass function of X, where X denotes the number of defects in a 1 m length of aluminium wire. One hundred wires are sampled from this population. Find the probability that the average number of defects per wire in this sample is less than 0.5.

X	0	1	2	3
P(X = x)	0.48	0.39	0.12	0.01

P3.95 Suppose that a random sample of size $n = 12$ is taken from the uniform distribution on the interval [0, 1]. Determine the value of $P\left(\left|\bar{X}_n - \frac{1}{2}\right| \le 0.1\right)$.

P3.96 Suppose that a random variable X has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2\ ; & 4 \le x \le 6 \\ 0\ ; & \text{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 50$. Use central limit theorem.

P3.97 At a large university, the mean age of the students is 21.3 years, and the standard deviation is 4 years. A random sample of 64 students is drawn. What is the probability that the average age of these students is greater than 22 years?

P3.98 The GPAs of all 5540 students enrolled at a university have an approximate normal distribution with a mean of 3.02 and a standard deviation of 0.29. Let $\bar{x}$ be the mean GPA of a random sample of 48 students selected from this university. Determine the mean and standard deviation of $\bar{x}$ and comment on the shape of its sampling distribution.

REVIEW QUESTIONS

1. Explain the meaning of probability distribution of a discrete random variable. What are the various ways to present the probability distribution of a discrete random variable?
2. Briefly explain the two-characteristics (conditions) of the probability distribution of a discrete random variable.
3. Define the terms mean and standard deviation of a discrete random variable.
4. Briefly explain the following:
   - *(a)* binomial experiment
   - *(b)* trial
   - *(c)* binomial random variable
5. Describe the parameters of the binomial probability distribution.
6. Define the following:
   - *(a)* *n*-factorial
   - *(b)* combinations
   - *(c)* permutations

7. What are the conditions that must be satisfied to apply Poisson probability distribution?
8. What is the parameter of the Poisson distribution?
9. Define the following terms:
    (*a*) Bernoulli trial
    (*b*) Binomial parameters
    (*c*) Binomial probability distribution
    (*d*) continuous random variable
    (*e*) discrete random variable
    (*f*) random variable
    (*g*) z-value or z-score
10. Describe the difference between the probability distribution of a discrete random variable and that of a continuous random variable.
11. Explain the main characteristics of a normal distribution.
12. Describe the standard normal distribution curve.
13. Describe the parameters of the normal distribution.
14. Describe the conditions for the normal distribution to be used as an approximation to the binomial distribution.

STATE TRUE OR FALSE

1. A random variable is a quantitative variable whose value depends on chance. (True/False)
2. A discrete random variable is a random variable whose possible values cannot be listed. (True/False)
3. Probability distribution is a listing of the possible values and corresponding probabilities of a discrete random variable, or a formula for the probabilities. (True/False)
4. Probability histogram is a graph of the probability distribution that displays the possible values of a discrete random variable on the horizontal axis and the probabilities of those values on the vertical axis. (True/False)
5. For any discrete random variable, $\Sigma P(X = 1) = 1$. (True/False)
6. The term expected value and expectation are commonly used in place of the term mean. (True/False)
7. $k! = k(k-1)\cdots 2{\cdot}1$. (True/False)
8. $0! = 1$. (True/False)
9. In Bernoulli trials, the trials are not independent. (True/False)
10. In Bernoulli trials, the probability of a success, called the success probability, remains the same from trial to trial. (True/False)
11. In Bernoulli trials, the experiment (each trial) has more than two possibilities. (True/False)
12. The binomial distribution is the probability distribution for the number of failures in a sequence of Bernoulli trials. (True/False)
13. In Bernoulli trials, the number of outcomes that contain exactly x successes equals the binomial coefficient $\begin{Bmatrix} n \\ x \end{Bmatrix}$. (True/False)

14. The mean of a binomial random variable with parameters n and p is $\mu = np$. (True/False)
15. The standard deviation of binomial random variable with parameters n and p is $\sigma = \sqrt{np(1-p)}$. (True/False)
16. The mean of a Poisson random variable with parameter λ is $\mu = \lambda$. (True/False)
17. The standard deviation of a Poisson random variable with parameter λ is $\sigma = \sqrt{\lambda}$. (True/False)
18. A random variable is a quantitative variable whose value depends on chance. (True/False)
19. A discrete random variable is a random variable whose possible values cannot be listed. (True/False)
20. The sum of the probabilities of the possible values of a discrete random variable equals 0. (True/False)
21. The number of possible permutations of m objects among themselves is $m!$. (True/False)
22. The number of possible samples of size n from a population of size N is NC_n. (True/False)
23. For any two events, the probability that one or the other of the events occurs equal the sum of the two individual probabilities. (True/False)
24. For any event, the probability that it occurs equals 1 minus the probability that it does not occur. (True/False)
25. Data obtained by observing values of one variable of a population are called univariate data. (True/False)
26. Data obtained by observing values of two variables of a population are called bivariate data. (True/False)
27. A frequency distribution for bivariate data is called contingency table, or two-way table. (True/False)
28. The joint probability equals the product of the marginal probabilities. (True/False)
29. For a normally distributed variable, the percentage of all possible observations within any specified range equals the corresponding area under its associated normal curve, expressed as a percentage. (True/False)
30. A normally distributed variable having mean 0 and standard deviation 1 is said to have the standard normal distribution. (True/False)
31. A normal distribution is completely determined by the mean and standard deviation. (True/False)
32. The shape of a normal distribution is completely determined by its standard deviation.(True/False)
33. The total area under the standard normal curve is –1. (True/False)
34. The standard normal curve extends indefinitely in both directions, approaching, but never touches, the horizontal axis as it does so. (True/False)
35. The standard normal curve is symmetric about 0. (True/False)
36. Almost all the area under the standard normal curve lies between –1 and +1. (True/False)
37. For a normally distributed variable, one can determine the percentage of all possible observations that lie within any specified range by first converting z-scores and then obtaining the corresponding area under the standard normal curve. (True/False)
38. The rule of thumb for using the normal approximation to the binomial is that np and $n(1-p)$ are 5 are greater. (True/False)

39. A variable is said to be normally distributed if its distribution has the shape of a normal curve. (True/False)
40. If a variable of a population is normally distributed and is the only variable under consideration, common practice is to say that the population is a normally distributed population. (True/False)
41. The parameters for a normal curve are the corresponding mean and standard deviation of the variable. (True/False)
42. Two variables that have the same mean and standard deviation have the same distribution. (True/False)
43. Two normally distributed variables that have the same mean and standard deviations have the same distribution. (True/False)
44. Two normal distributions that have the same mean are centred at the same place, regardless of the relationship between their standard deviations. (True/False)
45. Two normal distributions that have the same standard deviations have the same shape, regardless of the relationship between their means. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **True** 5. **True** 6. **True** 7. **True** 8. **True** 9. **False** 10. **True**
11. **False** 12. **False** 13. **True** 14. **True** 15. **True** 16. **True** 17. **True** 18. **True** 19. **True** 20. **False**
21. **True** 22. **True** 23. **False** 24. **True** 25. **True** 26. **True** 27. **True** 28. **True** 29. **True** 30. **True**
31. **True** 32. **True** 33. **False** 34. **True** 35. **True** 36. **False** 37. **True** 38. **True** 39. **True** 40. **True**
41. **True** 42. **False** 43. **True** 44. **True** 45. **True**

CHAPTER 4

Sampling Distributions

This chapter introduces the reader the basic sampling distributions. Inferential statistics is used here to draw conclusions about a population, based on a probabilistic model of random samples of the population. Since different random samples will most likely give different estimates, some knowledge of the variability of all possible estimates derived from random samples is needed to arrive at reasonable conclusions. *Population* is any finite set of objects being investigated. A sample of objects drawn from a population is a *random sample*.

If the populations contains N elements and a sample of n of them is to be selected, then if each of the $\frac{N!}{(N-n)!n!}$ possible samples has an equal probability of being chosen, the procedure employed is known as *random sampling*. Due to the difficulty in obtaining random samples in practice tables of random numbers are used.

Sampling theory is the study of relationships existing between a population and samples drawn from the population. Sampling theory is useful in finding whether observed differences between two samples are actually due to chance variation or whether they are really significant. A study of inferences made regarding a population by the use of samples drawn from it, along with indications of the accuracy of such inferences using probability theory is known as *statistical inferences*. A *statistic* is a function of the observations in a random sample, which is not dependent on unknown parameters. A *parameter* is in general an unknown constant. For example, the parameters of a normal distribution are μ and σ^2, whereas $\overline{X}$ and s^2 are statistics. The behaviour of sample statistics is needed in order to draw conclusions about a sample. The probabilistic distribution of a random variable defined on a space of random samples is called a *sampling distribution*. The behaviour of the sample statistics is described by a *sampling distribution*. Several sampling distributions are discussed in this chapter and their application to inferential statistics in Chapter 5 (Estimation) and Chapter 6 (Hypothesis Testing).

Suppose that $y_1, y_2, \ldots, y_n$ represents a sample. Then the *sample mean*

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{4.1}$$

and the *sample variance*

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} \tag{4.2}$$

are both statistics. These quantities are measures of the central tendency and dispersion of the sample, respectively. Sometimes $s = \sqrt{s^2}$, called the *sample standard deviation*, is used as a measure of dispersion.

4.1 PROPERTIES OF SAMPLE MEAN AND VARIANCE

The sample mean $\bar{y}$ is a point estimator of the population mean μ, and the sample variance s^2 is a *point estimator* of the population variance σ^2. An *estimator* of an unknown parameter is a statistic that corresponds to that parameter. Note that a point estimator is a random variable. A numerical value of an estimator computed from sample data, is called an *estimate*. There are several properties required of good *point estimators*. Two of the most important are the following:

1. The point estimator must be *unbiased*. The long-run average of expected value of the point estimator should be the parameter that is being estimated. An unbiasedness is a desirable property and this alone does not always make an estimator a good one.
2. An unbiased estimator must have *minimum variance*. The minimum variance point estimator has a variance that is smaller than the variance of any other estimator of that parameter. Here, we show that $\bar{y}$ and s^2 are unbiased estimators of μ and σ^2, respectively. Using the properties of expectation,

$$E(\bar{y}) = E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu \tag{4.3}$$

since the expected value of each observation y_i is μ. Thus, $\bar{y}$ is an unbiased estimator of μ.

Considering the sample variances s^2. We get

$$E(S^2) = E\left[\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] = \frac{1}{n-1}E(SS) \tag{4.4}$$

where $SS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *correct sum of squares* of the observation y_i. Now

$$E(SS) = E\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right] = E\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right]$$

$$\sum_{i=1}^{n}(\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2 / n) = (n-1)\sigma^2 \tag{4.5}$$

Hence,
$$E(S^2) = \frac{1}{n-1} E(SS) = \sigma^2 \tag{4.6}$$

Therefore, s^2 is an unbiased estimator of σ^2. The quantity $n - 1$ in Equation (4.5) is called the *number of degrees of freedom* of the sum of squares SS. If y is a random variable with variance σ^2 and $SS = \sum (y_i - \bar{y})^2$ has degrees of freedom, then

$$E\left(\frac{SS}{\nu}\right) = \sigma^2 \tag{4.7}$$

The number of degrees of freedom of a sum of squares is equal to the number of independent elements in that sum of squares. For example, $SS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ in Equation (4.5) consists of the sum of squares of the n elements $y_1 - \bar{y}$, $y_2 - \bar{y}$, ..., $y_n - \bar{y}$. These elements are not all independent since $SS = \sum_{i=1}^{n} (y_i - \bar{y}) = 0$; in fact, only $n - 1$ of them are independent, implying that SS has $n - 1$ degrees of freedom.

The *standard error* of a statistic is the standard deviation of its sampling distribution. If the standard error involves unknown parameters whose values can be estimated, substitution of these estimates into the standard error results in an *estimated standard error*. Suppose, we are sampling from a normal distribution with mean μ and variance σ^2. Now the distribution of $\bar{X}$ is normal with mean μ and variance σ^2/n, so the *standard error* of $\bar{X}$ is

$$\frac{\sigma}{\sqrt{n}} \tag{4.7a}$$

The *estimated standard error* of $\bar{X}$ is

$$\frac{s}{\sqrt{n}} \tag{4.7b}$$

4.2 POPULATION AND SAMPLING DISTRIBUTIONS

Here, we introduce the basic concept of population distribution and sampling distribution.

4.2.1 Population Distribution

The *population distribution* is the probability distribution of the population data. It is the probability distribution derived from the information on all elements of a population.

4.2.2 Sampling Distribution

A sample provides data for only a portion of an entire population and therefore we cannot expect the sample to yield perfectly accurate information about the population. The value of a population parameter is always constant. Any population data set has only one value of the population mean, μ. We would expect different samples of the same size drawn from the same population to give different values of the sample mean, $\bar{x}$.

The sample mean, $\bar{x}$, is therefore a *random variable* and it possesses a probability distribution. For each sample we can find a statistic, such as the mean, standard deviation, etc., which will vary from sample to sample. As a consequence, we obtain a distribution of the statistic which is called its *sampling distribution*. Sample statistics such as the mean, median, mode and standard deviation all possess sampling distributions. In general, the probability distribution of a sample statistic is called its sampling distribution. The probability distribution of $\bar{x}$ is called its sampling distribution. It lists the values that $\bar{x}$ can assume and the probability of each value of $\bar{x}$.

4.3 SAMPLING AND NONSAMPLING ERRORS

Since a sample provides data for only a portion of an entire population, we cannot expect the sample to give perfectly accurate information about the population. Hence, we can expect a certain information about the population. Hence, we can expect a certain amount of error called *sampling error* - will result simply because we are sampling. *Sampling error* is the difference between the value of a sample statistic and the value of the corresponding population parameter. In the case of the mean,

$$\text{Sampling error} = \bar{x} - \mu \tag{4.8}$$

assuming that the sample is random and no nonsampling error has been made.

A sampling error occurs due to chance. The errors that occur in the collection, recording and tabulation of data are called *nonsampling errors*. Such errors occur due to human mistakes and not chance. The larger the sample size, the smaller the sampling error tends to be in estimating a population mean, μ, by a sample mean $\bar{x}$.

Example E4.1

The mean ages of all students in a large university follow a distribution that is skewed to the right is 24 years and a standard deviation of 4 years. Find the probability that the mean age for a random sample of 36 students would be

(*a*) between 23 and 25 years

(*b*) less than 23 years.

SOLUTION:

Given the population mean = 24 years, and $n = 36$.

The standard deviation of the sample mean is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{36}} = 0.6667 \text{ years}$$

(*a*) For $\bar{x} = 23$: $z = \dfrac{23-24}{0.6667} = -1.50$

For $\bar{x} = 25$: $z = \dfrac{25-24}{0.6667} = 1.50$

$$P(23 < \bar{x} < 25) = P(-1.50 < z < 1.50) = (0.933193 - 0.066807) = 0.86639$$

(from the table in Appendix-*E*)

(*b*) For $\bar{x} = 23$: $z = \dfrac{23 - 24}{0.6667} = -1.50$

$P(\bar{x} < 25) = P(z < -1.50) = 0.066807$ (from the table in Appendix-*E*).

4.4 MEAN AND STANDARD DEVIATION OF $\bar{x}$

For a variable x and a given sample size, the distribution of the variable $\bar{x}$ is called *the sampling distribution of the sample mean.*

The mean and standard deviation of the sampling distribution of $\bar{x}$ are called the *mean and standard deviation of* $\bar{x}$ and are denoted by $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ respectively.

The standard deviation of $\bar{x}$ is also called the *standard error of* $\bar{x}$. There is a simple relationship between the mean of the variable $\bar{x}$ and the mean of the variable under consideration. They are equal, or $\mu_{\bar{x}} = \mu$. Thus, for any particular sample size, the mean of all possible samples means equals the populations mean. This equality holds true regardless of the size of the sample.

Hence, the mean of the sampling distribution of $\bar{x}$ is always equal to the mean of the population.

$$\mu_{\bar{x}} = \mu \tag{4.9}$$

The sample mean $\bar{x}$, is called an *estimate* of the population mean, μ. When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, that simple statistic is said to be an *unbiased estimator*. For the sample mean $\bar{x}$, $\mu_{\bar{x}} = \mu$. Therefore, $\bar{x}$ is an unbiased estimator of μ.

Standard Deviation of the Sample Mean

For samples of size n, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration divided by the square root of the sample size.

Hence, $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{4.10}$$

The above Eq. (4.10) assumes the sampling is done with replacement from a finite population or when it is done from an infinite population. When sampling is done without replacement from a finite population, the appropriate formula is

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\,\frac{\sigma}{\sqrt{n}} \tag{4.11}$$

where n denotes the sample size and N the population size.

The sample size is considered to be small compared to the population size if the sample is equal to or less than 5% of the population size.

That is, if $n/N \leq 0.05$.

When the sample size is small relative to the population size, there is little difference between sampling with and without replacement. In such cases, the two formulae Eqs. (4.10) and (4.11) for $\sigma_{\bar{x}}$ yield almost the same values. However, in most practical applications, the sample size is small relative to the population size.

The possible sample means cluster more closely around the population mean as the sample size increases, and therefore the larger the sample size, the smaller the sampling error tends to be in estimating a population mean by a sample mean.

The spread of the sampling distribution of $\bar{x}$ is smaller than the spread of the corresponding population distribution. In other words, $\sigma_{\bar{x}} < \sigma$. The standard deviation of the sampling distribution of $\bar{x}$ decreases as the sample size increases.

If the standard deviation of a sample statistic decreases as the sample size is increased, that statistic is said to be a *consistent estimator*. From Eq.(4.10), $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$, it is clear that as n increases, the value of $\sqrt{n}$ also increases and as a result, the value of $\dfrac{\sigma}{\sqrt{n}}$ decreases. Hence, the sample mean $\bar{x}$ is a consistent estimator of the population mean, μ.

Example E4.2

A population random variable X has mean 120 and standard deviation 15. Find the mean and standard deviation of the sample mean $\bar{X}$ for random samples of size 5 drawn with replacement.

SOLUTION:

For the population, $\mu = 120$, $\sigma = 15$.

The mean $\mu_{\bar{X}}$ and the standard deviation $\sigma_{\bar{X}}$ of $\bar{X}$ are given by

$$\mu_{\bar{X}} = \mu = 120$$

and

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{5}} = 6.7082$$

Example E4.3

A population random variable X has mean 120 and standard deviation 15. Find the mean and standard deviation of $\bar{X}$ if the population size is 300 and the samples of size 5 are drawn without replacement.

SOLUTION:

Here $N = 300$ and $n = 5$

Mean $\mu_{\bar{X}} = \mu = 120, \sigma = 15.$

and the standard deviation

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{15}{\sqrt{5}}\sqrt{\frac{300-5}{300-1}} = 6.6632$$

Example E4.4

Let $S = \{1, 3, 7, 9\}$. Find the probability distribution of the sample mean $\overline{x}$ for random samples of size 2 drawn with replacement.

SOLUTION:

Since S has 4 elements, there are $4^2 = 16$ random samples of size 2 drawn with replacement. These pairs and their average values are given in Table P4.4.

Table P4.4

Sample	$\overline{x}$	Sample	$\overline{x}$	Sample	$\overline{x}$	Sample	$\overline{x}$
(1, 1)	1	(1, 3)	2	(1, 7)	4	(1, 9)	5
(3, 1)	2	(3, 3)	3	(3, 7)	5	(3, 9)	6
(7, 1)	4	(7, 3)	5	(7, 7)	7	(7, 9)	8
(9, 1)	5	(9, 3)	6	(9, 7)	8	(9, 9)	9

The probability distribution of $\overline{x}$ is given in Table P4.4(a)

Table P4.4(a)

$\overline{x}$	1	2	3	4	5	6	7	8	9
$p(\overline{x})$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{4}{16}$	$\frac{2}{16}$	$\frac{1}{16}$	$\frac{2}{16}$	$\frac{1}{16}$

Example E4.5

Let $S = \{1, 3, 7, 9\}$

(*a*) list all samples of size 3, without replacement

(*b*) how many samples, without replacement, are there of size 4, size n?

SOLUTION:

(*a*) A sample of size 3, without replacement is a subset of S containing 3 elements. There are

$$\binom{4}{3} = 4$$

Subsets: $\{1, 3, 7\}, \{1, 3, 9\}, \{1, 7, 9\}, \{3, 7, 9\}$.

(*b*) For $n = 1, 2, 3, 4$, there are $\binom{4}{n}$ samples of size n; for $n > 4$, there are no samples of size n.

Example E4.6

Let $S = \{1, 3, 7, 9\}$. Find the probability distribution of the sample mean $\overline{x}$ for random samples of size 2 drawn without replacement.

SOLUTION:

Since S has 4 elements, there are $\binom{4}{2} = 6$ random samples of size 2 drawn without replacement. These, their average value, and the probability distribution of $\overline{x}$ are given in Tables P4.6 and P4.6(a).

Table P4.6

Sample	$\overline{x}$
(1, 3)	2
(1, 7)	4
(1, 9)	5
(3, 7)	5
(3, 9)	6
(7, 9)	8

Table P4.6(a)

$\overline{x}$	$p(\overline{x})$
2	1/5
4	1/6
5	2/6
6	1/6
8	1/6

Example E4.7

How many teams of 5 students can be randomly selected from a class of 10 female students and 15 male students? *(a)* how many teams will have all male students? *(b)* how many teams will have all female students? *(c)* how many teams will have 3 female students and 2 male students?

SOLUTION:

The number of 5-student teams is the number of ways at 5 students can be selected from a class of 25 students, or the number of samples of size 5 that can be selected, without replacement, from a population of size 25, which is

$$\binom{25}{5} = 53,130 \text{ (from the table in Appendix-}B\text{)}$$

(a) The number of teams that will have all male students is

$$\binom{15}{5} = 3003 \text{ (from the table in Appendix-}B\text{)}$$

(b) The number of teams that have all female students is

$$\binom{10}{5} = 252 \text{ (from the table in Appendix-}B\text{)}$$

(c) The number of teams that have 3 female students and 2 male students is

$$\binom{10}{3}\binom{15}{2} = (120)(150) = 12{,}600 \text{ (see Appendix-}B\text{).}$$

Example E4.8

Determine the most likely breakdown of male students and female students in a team of 5 randomly selected from 15 male and 10 female students.

SOLUTION:

The ratio of 15 male students to 10 female students is 3 to 2. Hence, a team of 3 male and 2 female students would occur at random. From Problem P4.7, we have

5 male students in a team = 3003

5 female students in a team = 252

3 female students and 2 male students in a team = 12600

In a similar manner, we obtain the following counts: (use the table in Appendix-*B*)

$$\text{1 male and 4 female students in a team} = \binom{15}{1}\binom{10}{4} = (15)(210) = 3150$$

$$\text{3 male and 2 female students in a team} = \binom{15}{3}\binom{10}{2} = (455)(45) = 20{,}475$$

$$\text{4 male and 1 female students in a team} = \binom{15}{4}\binom{10}{1} = (1365)(10) = 13{,}650$$

4.5 SHAPE OF THE SAMPLING DISTRIBUTION OF $\bar{x}$

In Section 4.4, we described the sampling distribution of the sample mean, that is, the distribution of the variable $\bar{x}$. It was shown there that the mean and standard deviation of $\bar{x}$ can be expressed in terms of the sample size and the population mean and standard deviation:

$$\mu_{\bar{x}} = \mu \quad \text{and} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

In this section, we describe the shape of the sampling distribution of $\bar{x}$ as related to the following two cases:

1. The population from which samples are drawn has a normal distribution.
2. The population from which samples are drawn does not have a normal distribution.

4.5.1 Sampling from a Normally Distributed Population

If the population variable x of a population is normally distributed with mean μ and standard deviation σ, then the sampling distribution of the sample mean, $\bar{x}$, will also be normally distributed with the mean μ and standard deviation $\sigma/\sqrt{n}$.

That is $\mu_{\bar{x}} = \mu$

and
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \qquad \frac{n}{N} \le 0.05 \tag{4.12}$$

From the sampling distribution of $\bar{x}$ for different sample sizes, n, it can be observed that the spread of the sampling distribution of $\bar{x}$ decreases as the sample size increases.

Example E4.9

The lengths of all machine parts made by a company have a distribution that is skewed to the right with a mean of 68 mm and a standard deviation of 4 mm. Find the probability that the mean length of a random sample of 100 parts produced by this company would be

(*a*) less than 67.8 mm

(*b*) between 67.5 mm and 68.7 mm

(*c*) within 0.6 mm of the population mean

(*d*) lower than the population mean by 0.5 mm or more.

SOLUTION:

Given $\mu = 68$ mm, $\sigma = 4$ mm and $n = 100$.

The standard deviation of the sample mean

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{100}} = 0.4 \text{ mm}$$

(*a*) For $\bar{x} = 67.8$ mm: $z = \dfrac{67.8 - 68}{0.4} = -0.50$

$P(\bar{x} < 67.8) = P(z < -0.50) = 0.308538$ (from the table in Appendix-*E*).

(*b*) For $\bar{x} = 67.5$: $z = \dfrac{67.5 - 68}{0.4} = -1.25$

For $\bar{x} = 68.7$: $z = \dfrac{68.7 - 68}{0.4} = 1.75$

$P(67.5 \le \bar{x} \le 67.8) = P(-1.25 \le z \le 1.75) = (-0.105650 + 0.959941) = 0.85429$

(from the table in Appendix-*E*).

(*c*) $P(\bar{x}$ within 0.6 mm of $\mu) = P(67.4 \le \bar{x} \le 68.6)$

For $\bar{x} = 67.4$: $z = \dfrac{67.4 - 68}{0.4} = -1.50$

For $\bar{x} = 68.6$: $z = \dfrac{68.6 - 68}{0.4} = 1.50$

$P(67.4 \le \bar{x} \le 68.6) = P(-1.50 \le z \le 1.50) = 0.933193 - 0.066807 = 0.86639$

(from the table in Appendix-*E*).

(*d*) $P(\bar{x}$ lower than μ by 0.5 mm or more$) = p(\bar{x} \le 67.5)$

$P(\bar{x} \le 67.5) = P(z \le -1.25) = 0.10565$ (from the table in Appendix-*E*).

For $\bar{x} = 67.5$: $z = \dfrac{67.5 - 68}{0.4} = -1.25$

4.5.2 Sampling from a Population that is not Normally Distributed

If the sampling is done from a population that is not normally distributed, then the shape of the sampling distribution of $\bar{x}$ is inferred from central limit theorem (see Chapter 3, Section 3.8).

4.6 APPLICATIONS OF THE SAMPLING DISTRIBUTION OF $\bar{x}$

It was shown in Chapter 3 (Section 3.8) on central limit theorem that for large samples, the sampling distribution of $\bar{x}$ is approximately normal, regardless of the distribution of the variable under considerations. The approximation becomes better with increasing sample size. In general, the farther the variable under consideration is from being normally distributed, the larger the sample size must be for a normal distribution to provide an adequate approximation to the distribution of $\bar{x}$. A sample size of 30 or more ($n \geq 30$) is large enough.

The sampling distribution of the sample mean can be summarised as follows:

If the variable x of a population has mean μ and standard deviation σ, then for samples of size n:

1. The mean of $\bar{x}$ equals the population mean, or $\mu_{\bar{x}} = \mu$.
2. The standard deviation of $\bar{x}$ equals the population standard deviation divided by the square root of the sample size, or $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$.
3. If x is normally distributed, so is $\bar{x}$, regardless of sample size.
4. If the sample size is large, $\bar{x}$ is approximately normally distributed, regardless of the distribution of x.

Example E4.10

Refer to Problem P4.2. Suppose the random variable X in Problem P4.2 is approximately normally distributed, determine $P(115 \leq \bar{X} \leq 125)$ for samples of size 5 drawn with replacement.

SOLUTION:

From the solution of P4.2, the mean and standard deviation of $\bar{X}$ are $\mu_{\bar{x}} = 120$ and $\sigma_{\bar{x}} = 6.7082$. $\bar{X}$ is approximately normally distributed. Hence,

$$P(115 \leq \bar{X} \leq 125) = P\left(\frac{115-120}{6.7082} \leq \frac{\bar{X}-120}{6.7082} \leq \frac{125-120}{6.7082}\right)$$

$$= P(-0.745 \leq z \leq 0.745) \qquad \text{(See the table in Appendix-}E\text{).}$$

Where z is the standard normal random variable.

$$= (-0.229650 + 0.770350) = 0.5407$$

Example E4.11

Refer to Problem P4.3. Suppose the random variable X in Problem P4.3 is approximately normally distributed, find $P(115 \leq \bar{X} \leq 125)$ for samples of size 5 drawn with replacement.

SOLUTION:

Refer to the solution of Problem P4.3. The mean and standard deviation of $\bar{X}$ are $\mu_{\bar{X}} = 120$ and $\sigma_{\bar{X}} = 6.66.2$. $\bar{X}$ is approximately normally distributed. Hence,

$$P(115 \le \bar{X} \le 125) = P\left(\frac{115-120}{6.6632} \le \frac{\bar{X}-120}{6.6632} \le \frac{125-120}{6.6632}\right) = P(-0.75039 \le z \le 0.75039)$$

where z is the standard normal random variable.

Using the standard normal table in Appendix-*E*, we have

$$P(-0.75039 \le z \le 0.75039) = (-0.226627 + 0.773373) = 0.54675$$

4.7 POPULATION AND SAMPLE PROPORTIONS

The *population proportion*, denoted by p, is obtained by taking the ratio of the number of elements in a population with a specific characteristic to the total number of elements in the population. In general, when we pick a sample,

$$\text{Sample proportion of an attribute} = \frac{\text{number of items in the sample having the attribute}}{\text{sample size}} \tag{4.13}$$

Hence, if X is the number of items having a certain attribute in a sample of size n, then *the sample proportion having the attribute* is the random variable X/n. The probability distribution of this statistic is called the *sampling distribution of the proportion.*

The *population* and *sample proportions*, denoted by p and $\hat{p}$, respectively, are calculated as

$$p = \frac{X}{N} \text{ and } \hat{p} = \frac{x}{n} \tag{4.14}$$

where N = total number of elements in the population

n = total number of elements in the sample

X = number of elements in the population that possess a specific characteristic

x = number of elements in the sample that possess a specific characteristic

The sampled population can be described by the following probability distribution:

x	Probability
0	1 – p
1	p

The computations for finding the mean and the variance of this population are shown in Table 4.1.

Table 4.1

Value x	Probability p(x)	xp(x)	$x^2p(x)$
0	1 – p	0(1 – p)	$0^2(1 - p)$
1	p	1p	1^2p
Σ	1	P	p

From column 3 in Table 4.1, $\Sigma x p(x) = p$.

Hence, the population mean μ is p or $\mu = p$ (4.15)

Similarly the sum in column 4 is $\Sigma x^2 p(x) = p$.

Hence, the variance of the population, σ^2 is given by

$\sigma^2 = \Sigma x^2 p(x) - \mu^2 = p - p^2 = p(1-p)$

Therefore, population standard deviation = $\sqrt{p(1-p)}$ (4.16)

4.8 SAMPLING DISTRIBUTION OF $\hat{p}$

Similar to the sample mean $\bar{x}$, the sample proportion, $\hat{p}$, is also a random variable. It possesses a probability distribution called its *sampling distribution*. It gives the various values that $\hat{p}$ can assume and their probabilities.

4.9 MEAN AND STANDARD DEVIATION OF $\hat{p}$

The *mean of* $\hat{p}$, which is the same as the mean of the sampling distribution $\hat{p}$, is always equal to the population proportion, p.

Hence, $\mu_{\hat{p}} = p$ (4.17)

The sample proportion, $\hat{p}$, is called an *estimator* of the population proportion, p. When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, that sample statistic is said to be an *unbiased estimator*. Since for the sample population, $\mu_{\hat{p}} = p$, $\hat{p}$ is an unbiased estimator of p. The *standard deviation of* $\hat{p}$, denoted by $\sigma_{\hat{p}}$, is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} \tag{4.18}$$

where p is the population proportion, $q = 1 - p$, and n is the sample size.

For large values of n ($n \geq 30$) the sampling distribution is very closely normally distributed. Note that the population is binomially distributed. The Eq. (4.18) is also valid for a finite population in which sampling is with replacement.

However, if $n/N \geq 0.05$, then $\sigma_{\hat{p}}$ is given by

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{N-n}{N-1}} \tag{4.19}$$

where the factor $\sqrt{\frac{N-n}{N-1}}$ is called the finite population correction factor.

In general, the sample size, n is small compared to the population size, N and Eq.(4.18) is used.

If the standard deviation of a sample statistic decreases as the sample size is increased, that statistic is called the *consistent estimator*. It is clear from Eq.(4.18) that as n increases, the value of $\sqrt{pq/n}$ decreases and the sample proportion, $\hat{p}$, is said to be the *consistent estimator* of the population proportion, p.

The shape of the sampling distribution of $\hat{p}$ is inferred from the central limit theorem described in Section 3.8 of Chapter 3. According to the central limit theorem, the *sampling distribution of* $\hat{p}$ is approximately normal for a sufficiently large sample size. In the case of a proportions, the sample size is considered to be sufficiently large if np and $n(1 - p)$ are both greater than 5. That is, if $np > 5$ and $n(1 - p) > 5$. This condition is the same as required for the application of the normal approximation to the binomial probability distribution described in Chapter 3 (Section 3.5).

Example E4.12

A company manufactures engine cylinders. The machine that is used to make these cylinders is known to produce 6% defective cylinders. If a sample of 100 cylinders are selected every week and inspected them for being good or defective. If 8% or more of the cylinders in the sample are defective, the process is stopped and the machine is readjusted. Determine the probability that based on a sample of 100 cylinders the process will be stopped to readjust the machine.

SOLUTION:

Here, $p = 0.06, q = 1 - p = 1 - 0.06 = 0.94$ and $n = 100$.

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.06(0.94)}{100}} = 0.023749$$

For $\hat{p} = 0.08$, we have

$$z = \frac{(0.08 - 0.06)}{0.023749} = 0.84$$

Therefore, $P(\hat{p} \geq 0.08) = P(z \geq 0.84) = P(1 - z \leq 0.84)$. Refer to the table in Appendix-*E*.

$P(\hat{p} \geq 0.08) = (1 - 0.799546) = 0.20045$

Example E4.13

Let m be the mean annual salary of faculty members in a college for 1995. Assume that the standard deviation of the salaries of these faculty members is \$50,000. What is the probability that the 1995 mean salary of a random sample of 100 faculty members was within \$10,000 of the population mean, m? Assume that $\frac{n}{N}$ £ 0.05.

SOLUTION:

Given $\sigma = \$50,000$ and $n = 100$

Standard deviation of the sample mean

$$\mu_{x} = \frac{\sigma}{\sqrt{n}} = \frac{50000}{\sqrt{100}} = \$5000$$

The required probability is

$P[\mu - 10,000 \leq \bar{x} \leq \mu + 10,000]$

For $\bar{x} = \mu - 10,000$: $z = [(\mu - 10,000) - \mu]/5,000 = -2$

For $\bar{x} = \mu + 10,000$: $z = [(\mu + 10,000) - \mu]/5,000 = 2$

Hence $P(\mu - 10{,}000 \le \bar{x} \le \mu + 10{,}000) = P(-2 \le z \le 2)$

$= (-0.022750 + 0.977250)$ from the table in Appendix-*E*

$= 0.9545$

Example E4.14

A particular city is planning to build a nuclear power plant to generate the electric power. An independent survey fund that 53% of the voters in that city favour the building of that plant. Assume that this result holds true for the population of all the voters in this city.

(*a*) what is the probability that more than 50% of the voters in a random sample of 200 voters selected from this city will favour the building of this plant?

(*b*) a city official would like to take a random sample of voters in which over 50% would favour the plant building. How large a sample should be selected so that the city official is 95% sure of this outcome? Assume $\frac{n}{N} \le 0.05$.

SOLUTION:

Given $p = 0.53$, $n = 200$ and $n/N = \le 0.05$

(*a*) Standard deviation of the sample mean

$$\sigma_{\bar{x}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.53)(0.47)}{200}} = 0.03529$$

The shape of the sampling distribution is approximately normal. In order to have over 50% in favour in a sample of 200 requires 101 or more in favour of the plant building. Hence, we require

$$P\left(\hat{p} \ge \frac{101}{200}\right) = P(\hat{p} \ge 0.505)$$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}} = \frac{0.505 - 0.53}{0.03529} = -0.71$$

$$P(\hat{p} \ge 0.505) = P(z \ge -0.71) = 1 - P(z \le -0.71) = 1 - 0.238852 = 0.76115$$

(from the table in Appendix-*E*).

(*b*) $P(z > -1.65) = 1 - P(z < -1.65) = 1 - 0.049471 = 0.95053$ (from the table in Appendix-*E*).

Since $$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

$$\sigma_{\hat{p}} = \frac{\hat{p} - p}{z} = \frac{0.5 - 0.53}{-1.65} = 0.01818$$

Since $$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$n = \frac{p(1-p)}{(\sigma_{\hat{p}})^2} = \frac{0.53(0.47)}{(0.01818)^2} = 753.53 \approx 754$$

Hence, the sample should include at least 754 voters.

Example E4.15

A sample of 49 is picked at random from a population of manufactured steel circular rods. If the standard deviation of the distribution of their diameters is known to be 3 cm, find the standard error of the mean if

(*a*) the population consists of 1000 steel rods

(*b*) the population is extremely large

SOLUTION:

(*a*) We are giving that $n = 49$, $N = 1000$, and $\sigma = 3$. Hence, the standard error of the mean is

$$\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3}{\sqrt{49}}\sqrt{\frac{1000-49}{1000-1}} = \frac{3}{7}(0.97568) = 0.418$$

(*b*) If the population size is extremely large (practically infinite), the standard error is given by

$$\frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{49}} = \frac{3}{7} = 0.4286.$$

4.10 THE CHI-SQUARE DISTRIBUTION

The *chi-square distribution* is sometimes written as χ^2 *distribution* and read as *chi-square distribution*. The values of a chi-square distribution are denoted by the symbol χ^2, just as the values of the standard normal distribution and the *t*-distribution are defined by *z* and *t*, respectively. A variable has a *chi-square distribution* if its distribution has the shape of a specific type of right-skewed curve, called a *chi-square* *(χ^2) curve*. There are infinitely many chi-square distributions. The chi-square curve is identified by its number of degrees of freedom.

Let $Z_1, Z_2, \ldots, Z_k$ be normally and independently distributed random variables, with mean $\mu = 0$ and variance $\sigma^2 = 1$. Then the random variable

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

has the probability density function

$$f_{\chi^2}(u) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} u^{(k/2)-1} e^{-u/2} \qquad u > 0$$

$$= 0 \qquad \text{otherwise} \qquad (4.20)$$

and is said to follow the chi-square distribution with *k* degrees of freedom abbreviated χ_k^2.

The mean and variance of the χ_k^2 distribution are

$$\mu = k \qquad (4.21)$$

and

$$\sigma^2 = 2k \qquad (4.22)$$

Several chi-square distributions are shown in Fig. 4.1.

Some basic properties of χ^2-curves are:

1. The total area under a χ^2-curve equals 1.
2. A χ^2-curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.
3. A χ^2-curve is right-skewed.
4. As the number of degrees of freedom becomes larger, χ^2-curves look increasingly like normal curves.

Percentages (and probabilities) for a variable having a chi-square distribution are equal to areas under its associated χ^2-curve.

The chi-square random variable is non-negative, and the probability distribution is skewed to the right. As k increases, the distribution becomes more symmetric. As $k \to \infty$, the limiting form of the chi-square distribution is the normal distribution. The percentage points of the χ^2_k distribution are given in Appendix-*F*. Define $\chi^2_{\alpha,k}$ as the percentage point or value of the chi-square random variable with k degrees of freedom such that the probability that χ^2_k exceed this value is α. That

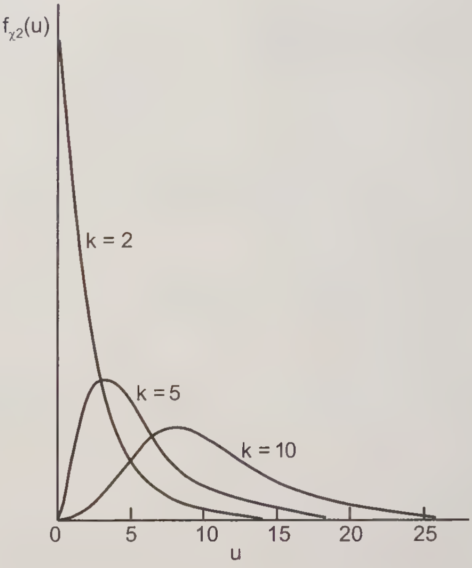


Fig. 4.1: Several χ^2 distributions

$$P\left\{\chi^2_k \geq \chi^2_{\alpha,k}\right\} = \int_{\chi^2_{\alpha,k}}^{\infty} f_{\chi^2}(\mu)\, du = \alpha \tag{4.23}$$

The probability in Eq. (4.23) is shown in Fig. 4.2.

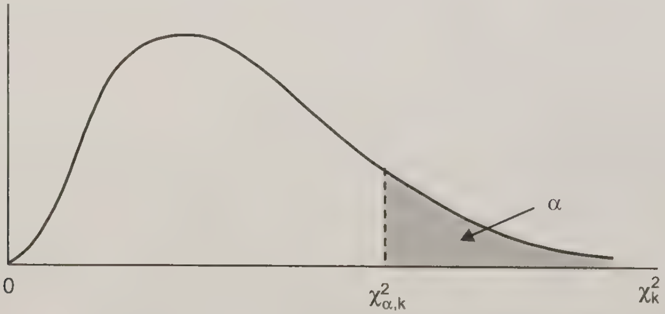


Fig. 4.2: Percentage point $\chi^2_{\alpha,k}$ of the chi-square distribution

The two outside columns of table in Appendix-*F*, labeled *df*, display the number of degrees of freedom. The symbol χ^2_α denotes the χ^2-value having area α to its right under a χ^2-value. Hence, the column headed $\chi^2_{0.995}$, for example, contains χ^2-values having area 0.995 to their right.

Example E4.16

Determine the χ^2-value having area 0.05 to the left for a χ^2-curve with $df = 9$ as shown in Fig. E4.16(*a*).

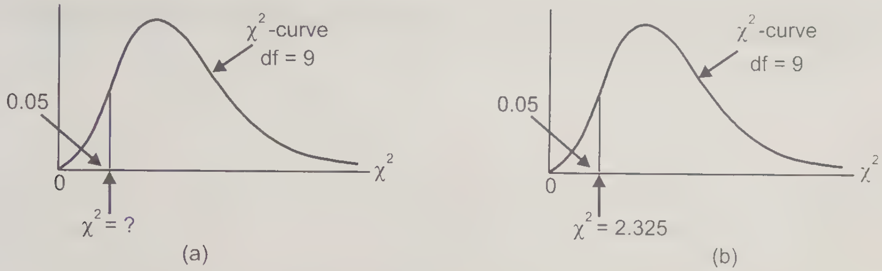


Fig. E4.16

SOLUTION:

The total area under a χ^2-curve equals $1 - 0.05 = 0.95$. Hence, the required χ^2-value is $\chi^2_{0.995}$. From the table in Appendix-*F* with $df = 9$, $\chi^2_{0.995} = 3.325$. Therefore, for a χ^2-curve with $df = 9$, the χ^2-values having area 0.05 to its left is 3.325 as shown in Fig. E4.16(*b*).

Example E4.17

For a χ^2-curve with $df = 25$, determine the two χ^2-values that divide the area under the curve into a middle 0.95 area and two outside 0.025 areas as shown in Fig. E4.17(*a*).

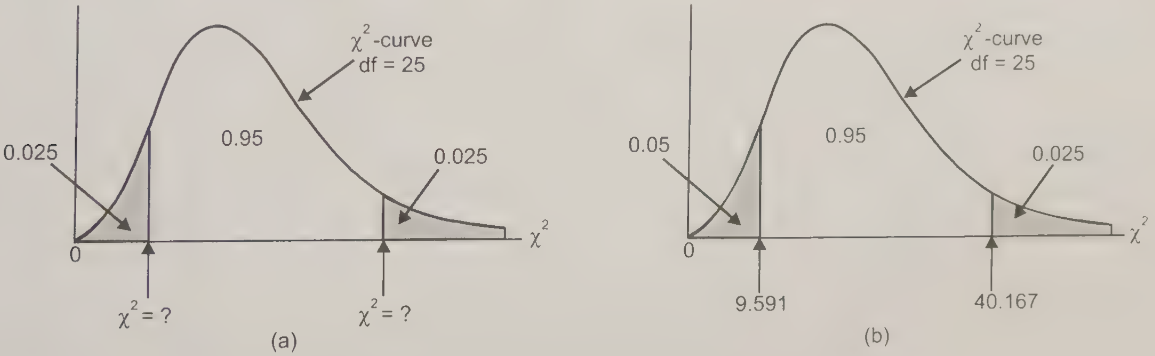


Fig. E4.17

SOLUTION:

First, we find the χ^2-value on the right in Fig. E4.17(*a*). Because the shaded area on the right is 0.025, the χ^2-value on the right is $\chi^2_{0.025}$. From the table in Appendix-*F* with $df = 25$, $\chi^2_{0.025} = 40.647$.

Next, we find the χ^2-value on the left in Fig. E4.17(*a*). Because the area to the left of that χ^2-value is 0.025, the area to its right is $1 - 0.025 = 0.975$. Hence, the χ^2-value on the left is $\chi^2_{0.975}$, which, by table in Appendix-*F* equals 13.120 for $df = 25$.

Consequently, for a χ^2-curve with $df = 25$, the two χ^2-values that the area under the curve into a middle 0.95 area and two outside 0.025 areas are 13.120 and 40.167, as shown in Fig. E4.17(*b*).

4.11 THE t-DISTRIBUTION

The standardised version of $\bar{x}$ has the standard normal distribution. William Gosset in 1908 determined the distribution of the studentised version of $\bar{x}$, a distribution now called *student's t-distribution* or, simply, the *t-distribution*.

There is a different *t*-distribution for each sample size. A particular *t*-distribution is identified by its number of *degrees of freedom* (*df*). For the studentised version of $\bar{x}$, the number of degrees of freedom is 1 less than the sample size, which is indicated symbolically by $df = n - 1$. The mean of the *t*-distribution is 0 just like the standard normal distribution. But unlike the standard normal distribution, whose standard deviation is 1, the standard deviation of a *t*-distribution is $\sqrt{\frac{df}{(df-2)}}$ which is always greater than 1. Hence, the standard deviation of a *t*-distribution is larger than the standard deviation of the standard normal distribution.

The number of degree of freedom is the only parameter of the *t*-distribution. A variable with a *t*-distribution has an associated curve, called a *t-curve*. There is a different *t*-distribution for each number of degrees of freedom.

The *t*-distribution is a specific type bell-shaped distribution with a height and a wider spread than the standard normal distribution. As the sample size becomes larger, the *t*-distribution approaches the standard normal distribution.

Let $Z \sim N(0, 1)$ and V be a chi-square random variable with k degrees of freedom. If Z and v are independent, then the random variable

$$\Gamma = \frac{Z}{\sqrt{V/k}} \tag{4.24}$$

has the probability density function

$$f(t) = \frac{\Gamma[(k+1)/2]}{\sqrt{\pi k}\,\Gamma(k/2)} \times \frac{1}{[(t^2/k)+1]^{(k+1)/2}} \qquad -\infty < t < \infty \tag{4.25}$$

and is said to follow the *t*-distribution with k degrees of freedom, abbreviated t_k.

The mean and variance of t are $\mu = 0$ and $\sigma^2 = k/(k-2)$ for $k > 2$, respectively. Several *t*-distributions are shown in Fig. 4.3.

The basic properties of *t*-curves are:

1. The total area under a *t*-curve equals 1.
2. A *t*-curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.
3. A *t*-curve is symmetric about 0.
4. As the number of degrees of freedom becomes larger, *t*-curves look increasingly like the standard normal curve.

Note that if $k = \infty$, the t-distribution becomes the standard normal distribution. A table of percentage points of the t-distribution is given in Appendix-*G*. If $x_1, x_2, \ldots, x_n$ is a random sample from $N(\mu, \sigma^2)$ distribution, then the quantity

$$t = \frac{x-\mu}{s/\sqrt{n}}$$

is distributed as t with $n - 1$ degrees of freedom.

We will let $t_{\alpha,k}$ be the value of the random variable Γ with k degrees of freedom above which we find an area (or probability) α. Thus $t_{\alpha,k}$ is an upper-tail 100α percentage point of the t-distribution with k degrees of freedom. This percentage point is shown in Fig. 4.4. In the Appendix-*G*, the α values are the column headings, and the degrees of freedom are listed in the left column. To illustrate the use of the table, note that the t-value with 15 degrees of freedom having an area of 0.05 to the right is $t_{0.05,10} = 1.753$. That is,

$$P(\Gamma_{10} > t_{0.05,15}) = P(\Gamma_{10} > 1.753) = 0.05$$

Since t-distribution is symmetric about zero, we have $t_{1-\alpha} = -t_{\alpha}$; that is, the t-value having an area of $1 - \alpha$ to the right (and therefore an area of α to the left) is equal to the negative of the t-value that has area α in the right tail of the distribution. Therefore, $t_{0.95,15} = -t_{0.05,15} = -1.753$.

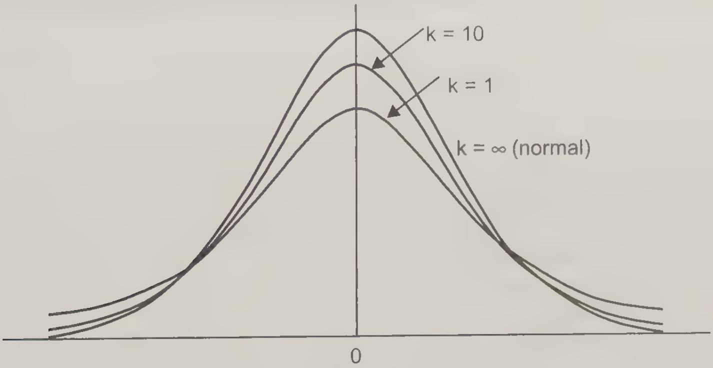


Fig. 4.3: Several *t*-distributions

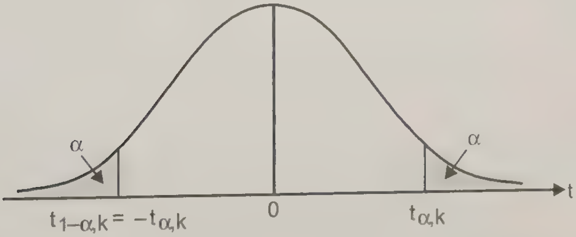


Fig. 4.4: Percentage points of the *t*-distribution

Example E4.18

Find a t-curve with 17 degrees of freedom, determine $t_{0.05}$. In other words, find the t-value having area 0.05 to its right, as shown in Fig. E4.18(*a*).

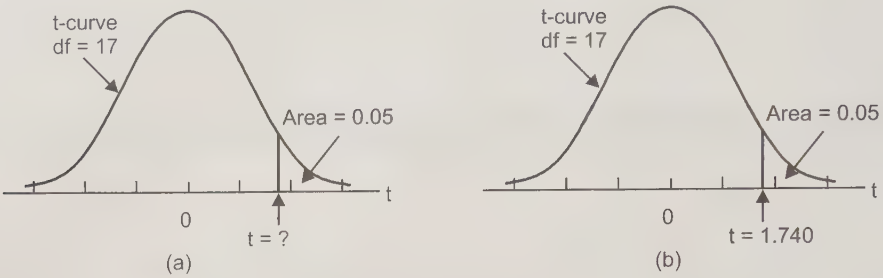


Fig. E4.18

SOLUTION:

The number degrees of freedom is 17, so we first go down the outside columns, (from the table in Appendix-*E*) labeled *df*, to "17". Then, going across that row to the column labeled $t_{0.05}$, we reach 1.740. This number is the *t*-value having area 0.05 to its right, as shown in Fig. E4.18(*b*). In other words, for a *t*-curve with $df = 17$, $t_{0.05} = 1.740$.

4.12 THE F-DISTRIBUTION

A variable is said to have an *F-distribution* if its distribution has the shape of a special type of right-skewed curve, called an *F-curve*. There are infinitely many *F*-distributions. The shape of a particular *F-distribution* curve depends on the number of degrees of freedom. There are two numbers of degrees of freedom for the *F*-distribution curve. The first number of degrees of freedom for an *F*-curve is called the *degrees of freedom for the numerator* and the second the *degrees of freedom for the denominator.*

The random variable *F* is defined to be the ratio of two independent chi-square random variables, each divided by its number of degrees freedom. That is

$$F = \frac{W/u}{Y/v} \tag{4.26}$$

where *W* and *Y* are independent chi-square random variables with *u* and *v* degrees of freedom, respectively.

Let *W* and *Y* be independent chi-square random variables with *u* and *v* degrees of freedom, respectively. Then the ratio

$$F = \frac{W/u}{Y/v} \tag{4.27}$$

has the probability density function

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}} \qquad 0 < x < \infty \tag{4.28}$$

and is said to follow the *F*-distribution with *u* degrees of freedom in the numerator and *v* degrees of freedom in the denominator. It is usually abbreviated as $F_{u,v}$.

The mean and variance of *F*-distribution are $\mu = v/(v-2)$ for $v > 2$, and

$$\sigma^2 = \frac{2v^2(u+v-2)}{u(v-2)^2(v-4)} \qquad v>4$$

The F random variable is non-negative, and the distribution is skewed to the right. The F-distribution looks very similar to the chi-square distribution as shown in Fig. 4.5; however the two parameters u and v provide extra flexibility regarding shape.

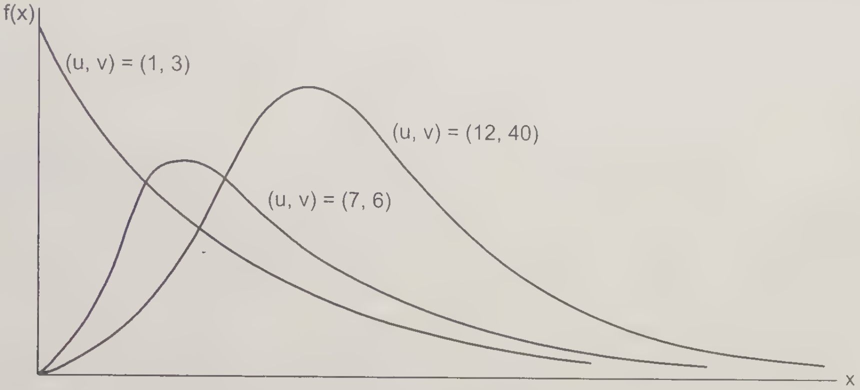


Fig. 4.5: Three *F*-distribution curves

The basic properties of F-curves are:

1. The total area under an F-curve equals 1.
2. An F-curve starts on 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so.
3. An F-curve is right-skewed.

For an F-curve with $df = (u, v)$, the F-value having area α to its left equals the reciprocal of the F-value having area α to its right for an F-curve with $df = (v, u)$.

The percentage points of F distribution are given in Appendix-H. Let $f_{\alpha,u,v}$ be the percentage point of F-distribution, with the numerator degrees of freedom u and the denominator degrees of freedom v such that the probability that the random variable F exceeds this value is

$$P(F > f_{\alpha,u,v}) = \int_{f_{\alpha,u,v}}^{\infty} f(x)\,dx = \alpha \tag{4.29}$$

This is illustrated in Fig. 4.6.

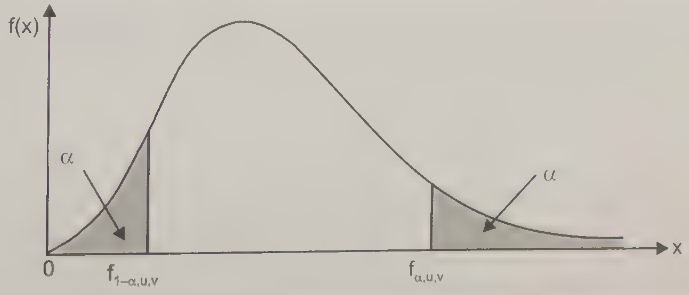


Fig. 4.6: Upper and lower percentage points of the F-distribution

For example, if $u = 6$ and $v = 10$, we find from the table in Appendix-*H*, that

$$P(F > f_{0.05,5,10}) = P(F_{5,10} > 3.22) = 0.05$$

That is, the upper 5 percentage point of $F_{5,10}$ is $f_{0.05,5,10} = 3.22$. The table in Appendix-*H* contains only upper-tail percentage points (for selected values of $f_{\alpha,u,v}$ for $\alpha \le 0.25$) of the *F*-distribution. The lower-tail percentage point $f_{1-\alpha,u,v}$ can be found as follows:

$$f_{1-\alpha,u,v} = \frac{1}{f_{\alpha,u,v}} \tag{4.30}$$

For example, to find the lower-tail percentage point of $f_{0.95,6,10}$, note that

$$f_{0.95,6,10} = \frac{1}{f_{0.05,10,6}} = \frac{1}{4.06} = 0.246$$

Example E4.19

For an *F*-curve with *df*(5, 13), find $F_{0.05}$. That is, find an *F*-value having area 0.05 to its right as shown in Fig. E4.19(*a*).

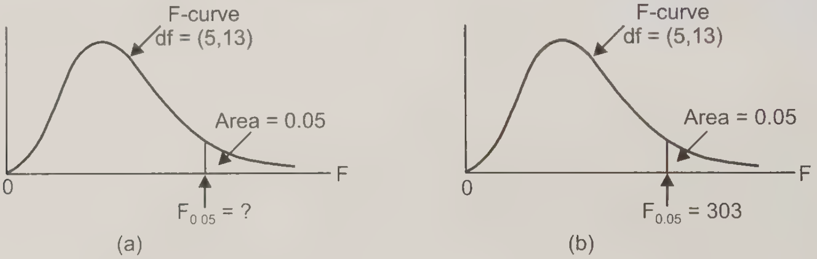


Fig. E4.19

SOLUTION:

To find the *F*-value, we use the table in Appendix-*H*. In this case, $\alpha = 0.05$ (area in the right tail under the *F*-distribution curve). The degrees of freedom for the numerator is 5, and the degrees of freedom for the denominator is 13.

We first go down the degree of freedom to "13". Next, we concentrate on the row for α labeled 0.05. Then, going across that row to the column labeled "5" (from the table in Appendix-*E*), we reach 3.03. This number is the *F*-value having area 0.05 to its right, as shown in Fig. E4.19(*b*). In other words, for an *F*-curve with *df*(5, 15), $F_{0.05} = 3.03$.

Example E4.20

For an *F*-curve with *df* = (50, 8) having an area 0.05 to its left.

SOLUTION:

The required *F*-value is the reciprocal of the *F*-value having area 0.05 to its right for an *f*-curve with *df* = (8, 50). From the table in Appendix-*H*, this latter *F*-value equals 2.13. Consequently, the required *F*-value is $\frac{1}{2.13}$ or 0.4695, as shown in Fig. E4.20.

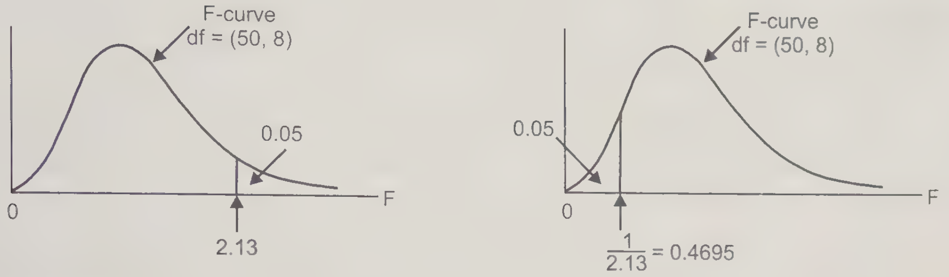


Fig. E4.20

Example E4.21

For an F-curve with $df = (10, 8)$, determine the two F-values that divide the area under the curve into a middle 0.95 area and two outside 0.025 areas as shown in Fig. E4.21(*a*).

SOLUTION:

First, we find the F-value on the right in Fig. E4.21(*a*). Because, the shaded area on the right is 0.025, the F-value on the right is $F_{0.025}$. From the table in Appendix-*H*, with $df = (10, 8)$, $F_{0.025} = 4.30$.

Next, we find the F-value on the left in Fig. E4.21(*a*). The F-value is the reciprocal of the F-value having area 0.025 to its right for an F-curve with $df(10, 8)$. From the table in Appendix-*H*, we find that this latter F-value equals $\frac{1}{4.30} = 0.2326$.

Consequently, for an F-curve with $df(10, 8)$, the two F-values that divide the area under the curve into a middle 0.95 area and two outside 0.025 areas are 0.2326 and 4.30, as shown in Fig. E4.21(*b*).

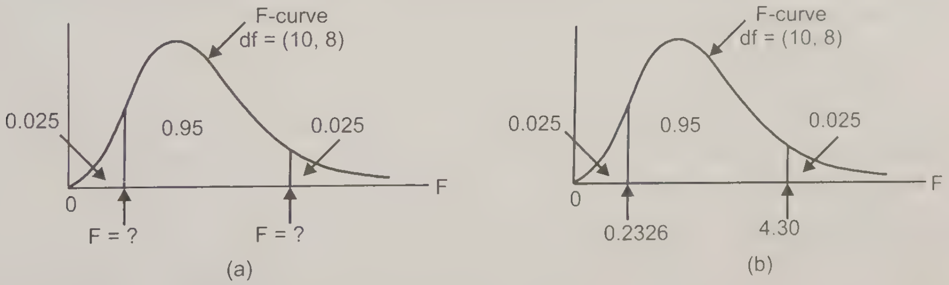


Fig. E4.21

4.13 SUMMARY

In Chapter 3, we discussed the probability distributions of discrete and continuous random variables. In this chapter, we extended the concept of probability distribution to that of a *sample statistic*. A sample statistic is a numerical summary measure calculated for sample data. The numerical summary measures calculated for population data are called *population parameters*. A population parameter is always a constant, whereas a sample statistic is always a random variable. Also, since every random variable must possess a probability distribution, each sample statistic possesses a probability distribution. The probability distribution of a sample statistic is more generally called its *sampling distribution*. In this chapter, we presented the sampling distributions of the sample mean and the sample proportion. The concepts presented in this chapter forms the foundation for the inferential statistics discussed in Chapters 5 and 6.

PROBLEMS

P4.1 A population consists of the values 1, 2 and 5.

(a) list all the possible samples (with replacement) of size $n = 2$ along with the sample means and their individual probabilities.

(b) find the mean of the sampling distribution

(c) does the sample mean target the values of the population mean?

P4.2 A manufacturing company produces machine parts that have a mean tensile strength of 100 MPa and a standard deviation of 10 MPa. The distribution of tensile strength is normal. Find the probability that a tensile strength less than 95 MPa.

P4.3 Suppose that a random variable X has a continuous uniform distribution

$$f(x) = \begin{cases} 1/2, & 4 \le x \le 6 \\ 0, & \text{otherwise} \end{cases}$$

Find the distribution of the sample mean of a random sample of size $n = 50$.

P4.4 Assume that the weights of 4000 male students at a university are normally distributed with a mean of 60 kg and standard deviation 3.0 kg. If 100 samples consisting 25 students each are obtained, what would be the expected mean and standard deviation of the resulting sampling distribution of means if sampling were done (*a*) with replacement, (*b*) without replacement?

P4.5 A survey showed that the mean expenditure incurred by a student in 1995 was \$9000 and the standard deviation of the expenditure was \$800. Find the approximate probability that the mean expenditure of 64 students picked at random was

(a) more than \$8820

(b) between \$8800 and \$9120

P4.6 Refer to Problem P4.4. In how many samples of Problem P4. 4 would you expect to find the mean

(a) between 58.8 and 60.3 kg

(b) less than 58.4 kg

P4.7 The length of life (in hours) of a certain type of machine component is a random variable with a mean of 600 hours and a standard deviation of 70 hours. What is the approximate probability that a random sample of 196 machine parts will have a mean life between 588 and 605 hours?

P4.8 Five hundred electronic components have a mean weight of 6.02 g and a standard deviation of 0.3 g. Find the probability that a random sample of 100 electronic components chosen from this group will have a combined weight of

(a) between 596 and 600 g

(b) more than 610 g

P4.9 Assume that the weights of machine parts of a heavy machine are normally distributed with mean 55 kg and standard deviation 2 kg. A number of samples of 100 parts each are taken at random with replacement from the population. Determine *(a)* the mean and standard deviation of the sampling distribution of the mean

(b) the probability that the sample mean will differ from the population mean by less than 0.4 kg.

P4.10 It was found in a particular survey that adults spend an average of 10 hours a day at work and commuting. Let the daily work and commute times for all adults have a mean of 10 hours and a standard deviation of 2 hours. Find the probability that the mean of the daily work and commute times for a random sample of 100 adults will be

(a) greater than 10.45 hours

(b) between 9.75 and 10.50 hours

(c) within 0.25 hours of the population mean

(d) lower than the population mean by 0.50 hours or more.

P4.11 A random sample of size 121 is taken from a normal population with $\sigma = 27.5$. What is the probability that the mean of the sample will differ from the mean of the population by 3 or more either way?

P4.12 The diameters of ball bearings manufactured by a company are normally distributed with mean 12 mm and standard deviation 0.1 mm. A sample of 25 ball bearings was taken each day during a given month. Find

(a) the measured standard deviation of the distribution of the sample means

(b) the probability that the sample mean of ball bearings will

  (i) exceed 12.01 mm

  (ii) be less than 11.98 mm

  (iii) lie between 11.98 and 12.1 mm

P4.13 A random sample of size of 81 is taken from a normal population with $\mu = 61.4$ and $\sigma = 7.65$. What is the probability that the mean of the sample will

(a) exceed 62.9

(b) fall between 60.5 and 62.3

(c) be less than 60.6

P4.14 A random sample of 300 air compressors had a standard deviation of 50 months. Determine the probability that the estimate on the population mean lifetime of these compressors will be within 6 months of the true mean from the sample estimate.

P4.15 According a particular survey, the standard deviation of the lengths of time that men with one job are employed during the first 10 years of their career is 100 weeks. Length of time employed during the first 10 years of career is left-skewed variable. For this variable, find the following:

(a) the sampling distribution of the sample mean for simple random samples of 100 men with one job

(b) the probability that the sampling error made in estimating the mean length of time employed by all men with one job by that of a random sample of 100 such men will be at most 25 weeks.

P4.16 In a large university, it was found that the score in statistics class with large number students over a number of years has a distribution with mean 80 and standard deviation 16. A random sample of 64 students of every class for 20 classes is taken. For the sampling distribution of mean, find (*a*) the mean and standard deviation

(*b*) the probability that

(*i*) $\bar{X} > 84$

(*ii*) $\bar{X} > 86$

(*iii*) $\bar{X} \leq 80$

where $\bar{X}$ is the sample mean found for the sample size of 64.

P4.17 According to a particular survey, the mean annual salary of private classroom teachers is \$45,000 and standard deviation of \$9000.

(*a*) find the sampling distribution of the sample mean for sample size of 81

(*b*) repeat part (*a*) for samples of 289

(*c*) is the assumption that classroom teachers salaries are normally distributed to answers in parts (*a*) and (*b*) necessary?

(*d*) what is the probability that the sampling error made in estimating the population mean salary of all classroom teachers by the mean salary of a sample of 81 classroom teachers will be at most \$1000

(*e*) repeat part (*d*) for samples of size 289.

P4.18 The weight of electronic components packed in certain containers is a random variable with a mean weight of 16 g and a standard deviation of 0.6 g. If the containers are shipped in boxes of 36, find, approximately, the probability that a randomly picked box will weigh over 585 g.

P4.19 According a particular survey it was found that the standard deviation of the lengths of hospital stay on the intervention ward is 9 days.

(*a*) for the variable "length of hospital stay", determine the sampling distribution of the sample mean for samples of 100 patients on the intervention ward

(*b*) the distribution of the length of hospital stay is right-skewed. Does this validate or invalidate the results found in part (*a*)?

(*c*) find the probability that the sampling error made in estimating the population mean length of stay on the intervention ward by the mean length of stay of a sample of 100 patients will be at most 3 days.

P4.20 The proportion of employed men between the ages 21 and 40 years in a city is 2/3. Suppose random samples of size 25 are drawn with replacement from all men in that city between the ages 21 and 40. What are the mean and standard deviation of the proportion $\hat{p}$ for all such samples?

P4.21 In a particular survey in a large metropolitan city nearly 20% of all students were at least somewhat afraid of being attacked in or nearby their schools. Suppose 20% of all students in grade 10 are afraid of such attacks. If $\hat{p}$ is the proportion of students in a random sample of 49 of grade 10 students who fear such attacks, find the mean and standard deviation of $\hat{p}$.

P4.22 Refer to Problem P4.20. Suppose the city has 250 men between ages 21 and 40 years, and the sampling is without replacement. What are the mean and standard deviation of $\hat{p}$?

P4.23 Suppose 25% of all workers in a state belong to a labour union. What is the probability that in a random sample of 100 workers in that state, at least 18% will belong to a labour union?

REVIEW QUESTIONS

1. Explain the meaning of a population distribution and a sampling distribution.
2. What is sampling error? Does such an error occur only in a sample survey or can it occur both in a sample survey and a census?
3. Explain the meaning of nonsampling error. Do such errors occur only in a sample survey or can they occur both in a sample survey and a census?
4. Describe the condition or conditions that must hold true for the sampling distribution of the sample mean to be normal when the sample size is less than 30.
5. Describe the central limit theorem (Refer to Chapter 3, Section 3.8).
6. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 2.5 standard deviations of the population mean? (Chapter 3, Section 3.9).
7. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 1.5 standard deviations of the population mean? (Chapter 3, Section 3.9).
8. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 2.0 standard deviations of the population mean? (Chapter 3, Section 3.9).
9. If all possible samples of the same (large) size are selected from a population, what percentage of all the sample means will be within 3.0 standard deviations of the population mean? (Chapter 3, Section 3.9).
10. Define the following terms:
    - *(a)* Consistent estimator
    - *(b)* Estimator
    - *(c)* Mean of $\hat{p}$
    - *(d)* Mean of $\overline{x}$
    - *(e)* Population proportion, p
    - *(f)* Unbiased estimator
11. Explain the following terms:
    - *(a)* Sample proportion
    - *(b)* Sampling distribution of $\hat{p}$
    - *(c)* Sampling distribution of $\overline{x}$
    - *(d)* Standard deviation of $\hat{p}$
    - *(e)* Standard deviation of $\overline{x}$
12. Describe very briefly the following distributions:
    - *(a)* t-distribution
    - *(b)* F-distribution
    - *(c)* chi-square distribution

STATE TRUE OR FALSE

1. Sampling error is the error resulting from using a sample to estimate a population characteristic. (True/False)
2. For a variable x and a given sample size, the distribution of the variable $\bar{x}$ is called the sampling distribution of the population mean. (True/False)
3. The larger the sample size, the smaller the sampling error tends to be in estimating a population mean, μ, by a sample mean $\bar{x}$. (True/False)
4. For a sample of size n, the mean of the variable $\bar{x}$ equals the mean of the variable under consideration. (True/False)
5. For samples of size n, the standard deviation of the variable $\bar{x}$ equals the standard deviation of the variable under consideration multiplied by the square root of the sample size. (True/False)
6. A normal distribution is determined by the mean and standard deviation. (True/False)
7. The mean of all possible sample means (i.e., of the variable $\bar{x}$) always equals the population mean. (True/False)
8. The larger the sample size, the larger is the standard deviation $\bar{x}$. (True/False)
9. The smaller the standard deviation of $\bar{x}$, the more closely the possible values $\bar{x}$ of (the possible sample means) cluster around the mean of $\bar{x}$. (True/False)
10. For a relatively large sample size, the variable $\bar{x}$ is approximately normally distributed.(True/False)
11. The total area under χ^2-curve equals 1. (True/False)
12. A χ^2-curve starts at 0 on the horizontal axis and extends indefinitely to the right, approaching, but never touching, the horizontal axis as it does so. (True/False)
13. A χ^2-curve is left skewed. (True/False)
14. As the number of degrees of freedom becomes larger, χ^2-curves look decreasingly like normal curves. (True/False)
15. The larger the standard deviation of $\bar{x}$, the more closely the possible values of $\bar{x}$ (the possible sample means) cluster around the mean of $\bar{x}$. (True/False)
16. For samples of size n, the variable $\chi^2 = \left(\frac{n-1}{\sigma^2}\right)s^2$ has the chi-square distribution with $n - 1$ degrees of freedom. (True/False)
17. A variable is said to have a chi-square distribution if its distribution has the shape of a special type of right-skewed curve, called a chi-square curve. (True/False)
18. Different chi-square distributions are identified by their degrees of freedom. (True/False)
19. The total area under an F-curve equals 1. (True/False)
20. An F-curve starts at 0, on the horizontal axis and extends indefinitely to the left, approaching, but never touching, the horizontal axis as it does so. (True/False)
21. An F-curve with $df = (v_1, v_2)$, the F-value having area α to its left equals the reciprocal of the F-value having area α to its right for an F-curve with $df = (v_2, v_1)$. (True/False)
22. An F-distribution and its corresponding F-curve are identified by stating its two numbers of degrees of freedom. (True/False)
23. An F-distribution has two numbers of degrees of freedom. (True/False)

24. The first degree of freedom in *F*-distribution is called the degrees of freedom for the numerator. (True/False)
25. The second degree of freedom in *F*-distribution is called the degrees of freedom for the denominator. (True/False)
26. A χ^2-curve looks increasingly like a normal curve as the number of degrees of freedom becomes larger. (True/False)
27. An *F*-curve is right-skewed. (True/False)
28. For an *F*-curve with $df = (15, 5)$, the *F*-value having area 0.05 to its left equals the reciprocal of the *F*-value having area 0.05 to its right for an *F*-curve with $df = (5, 15)$. (True/False)
29. The observed value of a variable having an *F*-distribution must be greater than or equal to 1. (True/False)
30. The total area under a *t*-curve equals 1. (True/False)
31. A *t*-curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so. (True/False)
32. A *t*-curve is not symmetric about 0. (True/False)
33. As a number of degrees of freedom becomes larger, *t*-curves look increasingly like the *F*-distribution curve. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **True** 5. **False** 6. **True** 7. **True** 8. **False** 9. **True** 10. **True**
11. **True** 12. **True** 13. **False** 14. **False** 15. **False** 16. **True** 17. **True** 18. **True** 19. **True** 20. **False**
21. **True** 22. **True** 23. **True** 24. **True** 25. **True** 26. **True** 27. **True** 28. **True** 29. **False** 30. **True**
31. **True** 32. **False** 33. **False**

CHAPTER 5

Estimation

Inferential statistics is the part of statistics that helps us to make decisions about some characteristics of a population based on sample information. *Estimation* is a procedure by which numerical value or values are assigned to a population parameter based on the information collected from a sample.

In inferential statistics, μ is called the *true population mean* and *p* is called the *true population proportion*. The value(s) assigned to a population parameter based on the value of a sample statistic is called an *estimate*. The sample statistic used to estimate a population parameter is called an *estimator*.

If the mean of the sampling distribution of a statistic equals to the corresponding population parameter, the statistic is called an *unbiased estimator* of the parameter, otherwise it is called a *biased estimator*. The corresponding values of such statistics are called *unbiased* or *biased estimates* respectively. If the sampling distributions of two statistics have the same mean (or expectation), the statistic with the smaller variance is called an *efficient estimator* of the mean while the other statistic is called an *inefficient estimator*. The corresponding values of the statistic are called *efficient* or *inefficient estimates* respectively.

Statistical inference is divided into two major categories: (1) Parameter estimation and (2) Hypothesis testing (see Chapter 6). Parameter estimation is further classified into two major types: (1) Point estimation and (2) Interval estimation.

5.1 POINT ESTIMATION

Estimation is a procedure by which numerical value or values are assigned to a population parameter based on the information collected from a sample. In inferential statistics, μ is called the *true population mean* and *p* is called the *true population proportion*. The value(s) assigned to a population parameter based on the value of a simple statistic is called an *estimate* of the population parameter.

If the population is small, we can ordinarily determine μ exactly by first taking a census and then computing μ from the population data. But if the population is large, as it often is in practice, taking a census is generally impractical, extremely expensive, or impossible. Nonetheless, we can usually obtain sufficiently accurate information about μ by taking a sample from the population.

An estimate of a population parameter given by a single number is called *point estimate* of parameter. An estimate of a population parameter is given by two numbers between which the parameter may be considered to lie is called an *interval estimator* of the parameter. The sample statistic used to estimate a population parameter is called an *estimator*. Hence, the sample mean, $\overline{X}$ is an estimator of the population mean, μ and the sample population, $\hat{p}$ is an estimator of the population proportion, p.

The desired estimators depend on the distribution used. These estimators are not always given by $\overline{X}$ and S^2. The common distributions and their estimators are given below.

Distribution	Parameter	Suggested Estimator(s)
Poisson	λ	$\hat{\lambda} = \overline{X}$
Binominal	p	$\hat{p} = \overline{X}$
Normal	μ, σ^2	$\hat{\sigma}^2 = S^2$ (denominator is $n - 1$)

After a sample is taken and the data are analysed, the result is an *estimate*.

The estimation procedure involves the following steps:

1. Select a sample
2. Collect the required information from the members of the sample
3. Calculate the value of the sample statistic
4. Assign value(s) to the corresponding population parameter

If we select a sample and compute the value of the sample statistic for this sample, the value gives the *point estimate* of the corresponding population parameter.

5.2 INTERVAL ESTIMATION

An interval estimate for a population parameter is called a *confidence interval*. In *interval estimation*, instead of assigning a single value to a population parameter, an interval is constructed around the point estimate and then a probabilistic statement that this interval contains the corresponding population parameter is made. This probabilistic statement is given by the *confidence level*. An interval that is constructed based on the confidence level is called a *confidence interval*. The confidence level associated with a confidence interval states how much confidence we have that the interval contains the true population parameter. The confidence level is denoted by $(1 - \alpha)100\%$. When expressed as probability, it is called the *confidence coefficient* and is denoted by $(1 - \alpha)$. Note that α is called the *significance level*.

A *tolerance interval* is another important type of interval estimate. For a normal distribution, we know that 95% of the distribution is in the interval $\mu - 1.96\ \sigma, \mu + 1.96\ \sigma$.

Confidence and tolerance intervals bound unknown elements of a distribution. A *prediction interval* provides bounds on one (or more) future observations from the population.

Summarising, the purpose of the 3 types of interval estimates are:

A confidence interval bounds population or distribution parameter (such as the mean weight of a person).

A tolerance interval bounds a selected proportion of a distribution.

A prediction interval bounds future observations from the population or distribution.

A point estimate of a parameter consists of a single value with no indication of the accuracy of the estimate. A confidence interval consists of an interval of numbers obtained from a point estimate of the parameter together with a percentage that specifies how confident we are that the parameter lies in the interval.

A confidence interval estimate of a parameter consists of an interval of number obtained from the point estimate of the parameter together with a 'confidence level' that specifies how confident we are that the interval contains the parameter. This is superior to a point estimate because it provides some information about the accuracy of the estimate whereas a point estimate does not.

Interval estimates indicate the precision or accuracy of an estimate and are therefore preferable to point estimates. A statement of the error of precision of an estimate is often called its *reliability*. An interval estimate consists of two sample statistics, L and U. The probability that the interval formed by L and U contains the true value of a parameter is $1 - \alpha$. For instance, to construct an interval estimate on the parameter θ, we find two statistics, L and U, such that

$$P(L \le \theta \le U) = 1 - \alpha \tag{5.1}$$

The resulting interval $[L \le \theta \le U]$ is called a $100(1 - \alpha)\%$ confidence interval on unknown parameter θ. If a large number of independent samples were taken from the population being considered, approximately $100(1 - \alpha)\%$ of the intervals formed would be expected to contain the true value of θ. Confidence intervals that contain both an L, or lower side, and a U, or upper side, are called *two-sided confidence intervals*. Confidence intervals that contain only a lower or an upper side are called *one-sided confidence intervals*. A one-sided $100(1 - \alpha)\%$ confidence interval on θ is given by

$$L \le \theta$$

with the probability property

$$P(L \le \theta) = 1 - \alpha \tag{5.2}$$

Similarly, the one-sided upper $100(1 - \alpha)\%$ confidence interval on θ is given by $\theta \le U$ with the probability property.

$$P(\theta \le U) = 1 - \alpha \tag{5.3}$$

The length of the observed confidence interval is an important measure of the quality of the information obtained from the sample. The half-interval length $\theta - L$ or $U - \theta$ is called the *accuracy* of estimator. The longer the confidence interval, the more confident we are that interval actually contains the true value of θ and the less information we have about the true value of θ.

Example E5.1

The following readings give the weights of 6 persons (in kg) picked at random from college students in a particular college: 50, 52, 55, 60, 65 and 70. Find estimates of the following:

(*a*) the true (population) mean weight of all the students

(*b*) the true variance of weights of all the students

(*c*) the true standard deviation of weights of all the students.

SOLUTION:

$$n = 6$$

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{352}{6} = 58.6667 \text{ kg}$$

$$\Sigma(X_i - \bar{X})^2 = (50 - 58.6667)^2 + (52 - 58.6667)^2 + \cdots = 303.3333$$

$$\text{Variance} = \frac{\Sigma(X_i - X)^2}{n-1} = \frac{303.3333}{6-1} = 60.6667 \text{ kg}^2$$

$$\text{Standard deviation} = \sqrt{\text{variance}} = \sqrt{60.6667} = 7.7889 \text{ kg}$$

(a) The estimate of the true mean weight is $\bar{X}$, that is, 58.6667 kg

(b) The estimate (in kg²) of the true variance is

$$\frac{\Sigma(X_i - \bar{X})^2}{n-1} = 60.6667$$

(c) The estimate (in kg) of the true standard deviation is $\sqrt{60.6667} = 7.7889$ kg

Example E5.2

A simple random sample of 44 engineers in New York city yielded the following data on their monthly income from employment (in thousands of dollars):

8	14	8	11	7	4	5	7	5
12	6	7	7	4	8	12	7	9
12	6	9	6	5	7	3	10	10
12	8	8	10	5	6	3	8	11
4	5	7	6	11	7	6	8	

(a) use the data to obtain a point estimate for the mean monthly income of all engineers working in New York city

(b) is your point estimate in part *(a)* likely to equal μ exactly?

SOLUTION:

(a) $\Sigma X_i = 334$, and $n = 44$

Hence $\bar{X} = 334/44 = 7.6$

(b) It is not likely that $\bar{X}$ is exactly equal to μ (population mean). Some sampling error is to be expected.

5.3 CONFIDENCE INTERVAL ON MEAN, VARIANCE KNOWN

This section presents methods for using sample data to find a point estimate and confidence interval estimate of a population mean. The key requirement in this section is that in addition to having sample data, we also know σ, the standard deviation of the population.

For obtaining a confidence interval for μ, in this section, we make the following assumptions:

1. The sample consists of n independent observations, that is, any one observation does not influence any other.
2. The underlying population is normally distributed with mean μ, which is, of course, unknown ($n > 30$).
3. The population standard deviation σ is known.

Let X be a random variable with unknown mean μ and known variance σ^2, and suppose that a random sample of size n, $X_1, X_2, \ldots, X_n$ is taken. A 100(1 – α)% confidence interval on μ can be obtained by considering the sampling distribution of the sample mean $\bar{X}$. The mean of $\bar{X}$ is μ and the variance σ^2/n. Therefore, the distribution of the statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is taken to be a standard normal distribution.

The distribution of $Z = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ is shown in Fig. 5.1. From this figure we note that

$$P\{-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\} = 1 - \alpha$$

or $$P\left\{-Z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}\right\} = 1 - \alpha$$

This can be arranged as

$$P\left\{\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}\right\} = 1 - \alpha \tag{5.4}$$

From Eqs. (5.1) and (5.4), we note that the 100(1 – α)% two-sided confidence interval on μ is

$$\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} \le \mu \le \bar{X} + z_{\alpha/2}\sigma/\sqrt{n} \tag{5.5}$$

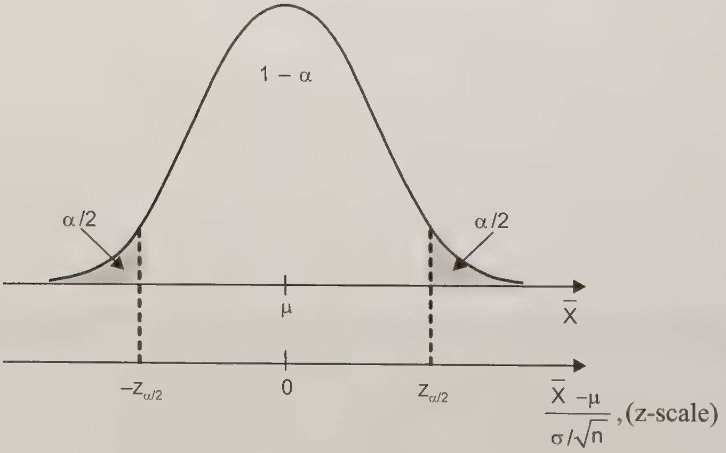


Fig. 5.1: The distribution of z

From Eq. (5.4), we observe that $\bar{X}$ will differ from μ by at most $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ with probability (1 – α). The quantity $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is, therefore, called the *maximum error of estimate* of μ at the (1 – α)100 per cent level. The maximum error of estimate of μ is commonly called the *margin of error* or the *sampling error.*

Hence the margin of error of estimate of μ at the $(1 - \alpha)100$ per cent level is the maximum error in estimating μ and is given by

$$\text{margin of error} = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The interval $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$ is a random interval and the probability that it covers the mean μ is $(1 - \alpha)$. The end points

$$L = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

and

$$U = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

of the interval are random variables. Their values depend on the value of $\bar{X}$ which, in turn, depends on the sample values.

Therefore, a confidence interval for μ is often given simply by specifying the limits as

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $\bar{X}$ is the sample mean based on n observations.

The above construction of the confidence interval for μ was based on the assumption that the population is normally distributed. This assumption was necessary since it permitted us to proceed by stating that $\bar{X}\left(\text{and } \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)$ has a normal distribution. If n is large (at least 30), the assumption of a normal distribution is crucial, because the central limit theorem allows us to proceed by stating that $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ has approximately a normal distribution (we still get the same confidence limits given above, but now they are approximate limits). When we say that the confidence interval is *exact*, we mean that the true confidence level equals $1 - \alpha$; similarly, when we say that the confidence interval is *approximately correct*, we mean that the true confidence level only approximately equals $1 - \alpha$.

The procedure to find a confidence interval for a population mean when the standard deviation is known is shown in Table 5.1.

Example E5.3

Refer to Example E5.1. Construct a 90% confidence interval for μ, the true mean weight of the students.

SOLUTION:

From Example E5.1, we have

$$n = 6,\ \bar{X} = 58.6667 \text{ kg and } \sigma = 7.7889 \text{ kg}.$$

For a 90% confidence interval, $\alpha = 0.10$. Thus, $z_{\alpha/2} = z_{0.05} = 1.645$ (from the table in Appendix-*E*). The confidence interval is given by

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$58.6667 - 1.645\frac{7.7889}{\sqrt{6}} \leq \mu \leq 58.6667 + 1.645\frac{7.7889}{\sqrt{6}}$$

or

$$58.6667 - 5.2308 \leq \mu \leq 58.6667 + 5.2308$$

$$53.4359 \leq \mu \leq 63.8975$$

which can be simplified to $53.44 < \mu < 63.90$, giving the confidence interval (53.44, 63.90). Hence, with 90% confidence we estimate that the mean weight μ of the students in that college is between 53.44 and 63.90 kg.

The term *normal population* is used here as an abbreviation for "the variable under consideration is normally distributed". The z-interval procedure works reasonably well even when the variable is not normally distributed and the sample size is small, or moderate, provided the variable is not too far from being normally distributed. For large samples, say, of size 30 or more ($n \geq 30$), the z-procedure can be used essentially without restriction. For samples of moderate size ($15 \leq n \leq 30$), the z-interval procedure can be used unless the data contains outliers or the variable under consideration is far from being normally distributed. For small samples, say, of size less than 15 ($n \leq 15$), the z-interval procedure should be used only when the variable under consideration is normally distributed or very close to being so. If outliers are present then their removal is justified in a data set in order to use the z-interval procedure.

Table 5.1: Procedure to find a confidence interval for a population mean when σ is known

Assumptions:

1. Simple random sample
2. Normal population or large sample
3. σ known

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*E* to find $z_{\alpha/2}$.

Step 2: The confidence interval for μ is from

$$\bar{X} - z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is found in step 1, n is the sample size and $\bar{X}$ is computed from the sample data.

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E5.4

Refer to Example E5.2. Assume that the recent monthly income of engineers from employment in New York city are normally distributed with a standard deviation of $8100.

(a) determine a 95.44% confidence interval for the mean cost, μ, of all recent engineers in New York city

(b) interpret your result in part *(a)*

(c) does the mean monthly income from employment of all engineers in New York city lie in the confidence interval obtained in part *(a)*?

Explain your answer.

SOLUTION:

(a) The confidence interval is given by

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

where $x_{\alpha/2} = 2.0$ (from the table in Appendix-*E*).

Therefore $7.6 - 2\frac{(2.4)}{\sqrt{44}}$ to $7.6 + 2\frac{(2.4)}{\sqrt{44}}$

or 6.9 to 8.3

(b) Since we know that 95.44% of all samples of 44 engineers employment monthly income have the property that the interval from $\bar{X} - 0.7$ to $\bar{X} + 0.7$ contains μ, we can be 95.44% confident that the interval from 6.9 to 8.3 contains μ.

(c) The confidence interval in part *(a)* would be exact if the population of engineers` monthly employment income were exactly a normal distribution. However, since engineers` monthly employment income is a discrete random variable and the normal distribution is continuous, monthly employment income cannot follow a normal distribution exactly.

Example E5.5

It has been established that in a certain packaging process, the packages are of an average weight of μ but that μ changes over time as the process adjustment changes through wear. It was also found that the variance of the package weight is a constant at 16 kg even though the mean changes overtime. A sample of 36 packages has revealed their mean weight $\bar{X}$ of 100 kg. Assuming that the weight of the individual packages are normally distributed around μ, *(a)* find the 90% confidence limits for μ, *(b)* find the 95% confidence limits for μ.

SOLUTION:

(a) Here $1 - \alpha = 0.90$ or $\alpha = 0.10$ and $\alpha/2 = 0.05$.

$z_{0.05} = 1.645$ (from the table in Appendix-*E*), $\alpha = \sqrt{16} = 4$ kg

By computation,

$$L = \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 - \frac{1.645(4)}{\sqrt{36}} = 100 - 1.0967 = 98.9033$$

$$U = \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 + \frac{1.645(4)}{\sqrt{36}} = 100 + 1.0967 = 101.0967$$

Hence, we are 90% confident that m lies between 98.9033 and 101.0967 kg.

(b) Here $z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*)

$$L = \bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 - \frac{1.96(4)}{\sqrt{36}} = 100 - 1.3067 = 98.6933$$

$$U = \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}} = 100 + \frac{1.96(4)}{\sqrt{36}} = 100 + 1.3067 = 101.3067$$

Therefore, we are 95% confident that μ lies between 98.6933 and 101.3067 kg.

One-sided Confidence Intervals

One-sided confidence intervals for μ are obtained by setting either $L = -\infty$ or $U = \infty$ and replacing $z_{\alpha/2}$ by z_α. The 100(1 – α)% upper-confidence interval for μ is

$$\mu \le \bar{X} + z_\alpha \sigma / \sqrt{n} \tag{5.6}$$

and the 100(1 – α)% lower-confidence interval for μ is

$$\bar{X} - z_\alpha \sigma / \sqrt{n} \le \mu \tag{5.7}$$

It is most preferable to obtain and pinpoint the parameter with 100% certainty. Due to the variability inherent in the population, it is hard to accomplish this. One can try to achieve the turn goal of a high level of confidence and a narrow interval.

Length of the confidence interval for μ: The length of any interval from *a* to *b* is (*b* – *a*). Hence, the length of the confidence interval is

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$$\left(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) - \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Therefore, the length of the confidence interval for μ

$$= 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 2\ \text{(margin of error)}$$

The length of the confidence interval is also called its *width*.

Hence, the length of the confidence interval is twice the margin of error or the margin of error is one-half the length of the confidence interval.

We also note that the length of the confidence interval for μ does not depend on $\bar{X}$, but it depends on $z_{\alpha/2}$, σ, and *n*. Hence, in order to achieve a high accuracy in estimating μ (that is, a narrow confidence interval with high degree of confidence) then one way to accomplish this goal is to pick an appropriately large sample. The margin of error is the standard error of the mean multiplied by $z_{\alpha/2}$. The length of a confidence interval, and thus the precision with which $\bar{X}$ estimates μ, is determined by the margin of error. Increasing the confidence level while keeping the sample size the same will increase the value of $z_{\alpha/2}$ and hence the length of the confidence interval (decrease the precision of the estimate). Increasing the sample size while keeping the same confidence level will decrease the margin of error and the length of the confidence interval (increase the precision).

Example E5.6

A random sample of 18 venture-capital investments in the fiber optics sector yields the following data, in crores of rupees:

2.04	5.48	5.60	5.96	6.27	10.51
4.13	5.58	5.74	5.95	6.67	8.63
4.21	4.98	6.66	7.71	8.64	9.21

(a) determine a 95% confidence interval for the mean amount, μ, of all venture, capital investments in the fiber optics sector. Assume that the population standard deviation is 2.04 crores of rupees

(b) interpret your finding in part *(a)*

(c) find a 99% confidence interval for μ

(d) why is the confidence interval found in part (*c*) longer than the one in part *(a)*?

(e) which confidence interval yields a more precise estimate of μ? Explain your answer.

SOLUTION:

(a)

$$n = 18 \text{ and } \Sigma X = 113.97 \text{ crores}$$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{113.97}{18} = 6.33 \text{ crores}$$

The 95% confidence interval for μ is

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$z_{\alpha/2} = 1.96 \text{ (from the table in Appendix-}E\text{)},$$

$$6.33 - 1.96\frac{(2.04)}{\sqrt{18}} \text{ to } 6.33 + 1.96\frac{(2.04)}{\sqrt{18}}$$

or 5.39 to 7.27 crores.

(b) We can be 95% confident that the interval from 5.39 crores to 7.27 crores contains the population mean venture capital investment in the fiber-optics sector.

(c) $n = 18$, $\bar{X} = 6.33$ and $\sigma = 2.04$

$\alpha = 0.01$ and $z_{\alpha/2} = z_{0.005} = 2.575$ (from the table in Appendix-*E*),

Hence

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ to } \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$6.33 - 2.575\frac{(2.04)}{\sqrt{18}} \text{ to } 6.33 + 2.575\frac{(2.04)}{\sqrt{18}}$$

or 5.09 to 7.57 crores.

(d) The confidence interval in *(c)* is longer than the one part *(a)* because we have changed the confidence level from 95% in part *(a)* to 99% in part *(c)*. We notice that increasing the confidence level from 95% to 99% increases the $z_{\alpha/2}$ value from 1.96 to 2.575. The larger z-value, in turn, results in a longer interval. In order to accomplish a higher level of confidence that the interval contains the population mean, we need a longer interval.

(e) See Fig. E5.6.

The 95% confidence interval is shorter and therefore provides a more precise estimate of μ.

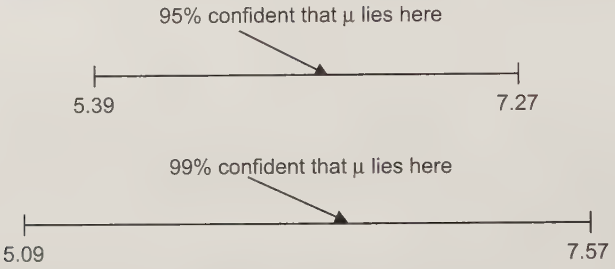

95% confident that μ lies here
5.39
7.27
99% confident that μ lies here
5.09
7.57

Fig. E5.6

5.4 CONFIDENCE INTERVAL ON THE MEAN OF A NORMAL DISTRIBUTION, VARIANCE UNKNOWN

The method presented here is valid for any arbitrating sample size but is particularly important when the sample size is small. For setting a confidence interval for μ, we make the following assumptions:

1. The sample is a simple random sample.
2. The observations are picked from the population under study and are independent.
3. The population has a normal distribution.

Here we find a confidence interval on the mean of a distribution when the variance is unknown. Suppose a random sample of size n, $X_1, X_2, \ldots, X_n$ is available and $\bar{x}$ and S^2 are the sample mean and sample variance, respectively. In order to use a valid confidence interval when the sample size small, we make an assumption that the underlying population is normally distributed. This leads to confidence intervals based on the t-distribution. Therefore, let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with unknown mean μ and unknown variance σ^2. We know the sampling distribution of the statistic

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

is the t-distribution with $n - 1$ degrees of freedom.

The distribution of $t = (\bar{X} - \mu)/(S/\sqrt{n})$ is shown in Fig. 5.2. Letting $t_{\alpha/2,\,n-1}$ be the upper $\alpha/2$ percentage point of the t-distribution with $n - 1$ degree of freedom, we note from Fig. 5.2, that

$$P\left\{-t_{\alpha/2,\,n-1} \le t \le t_{\alpha/2,\,n-1}\right\} = 1 - \alpha$$

or

$$P\left\{-t_{\alpha/2,\,n-1} \le \frac{\bar{X}-\mu}{S/\sqrt{n}} \le t_{\alpha/2,\,n-1}\right\} = 1-\alpha \tag{5.8}$$

Rearranging Eq. (5.8)

$$P\left\{\bar{X} - t_{\alpha/2,\,n-1} S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,\,n-1} S/\sqrt{n}\right\} = 1-\alpha \tag{5.9}$$

Comparing Eqs. (5.4) and (5.9), we see that a 100 (1 – α)% two-sided confidence interval on μ is

$$\bar{X} - t_{\alpha/2,\,n-1} S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,\,n-1} S/\sqrt{n} \tag{5.10}$$

A 100 (1 – α) % lower-confidence interval on μ is given by

$$\bar{X} - t_{\alpha/2,\,n-1} S/\sqrt{n} \le \mu \tag{5.11}$$

and a 100 (1 – α) % upper-confidence interval on μ is given by

$$\mu \le \bar{X} + t_{\alpha/2,\,n-1} S/\sqrt{n} \tag{5.12}$$

The above method assumes that the samplings are from a normal population.

The procedure to find a confidence interval for a population mean when the standard deviation σ is unknown, is given in Table 5.2. Properties and guidelines for use of the *t*-interval procedure are the same as those for the *z*-interval procedure. The *t*-interval procedure is robust to moderate violations of the normality assumptions.

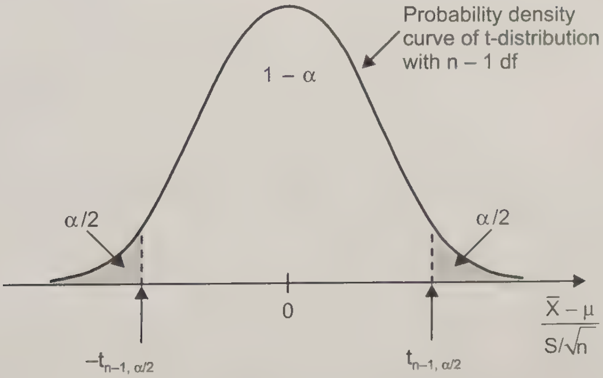

Probability density curve of t-distribution with n – 1 df
1 – α
α/2
α/2
0
$\frac{\bar{X}-\mu}{S/\sqrt{n}}$
$-t_{n-1,\,\alpha/2}$
$t_{n-1,\,\alpha/2}$

Fig. 5.2: t-values such that there is an area α/2 in the right tail and an area α/2 in the left tail of the distribution

Table 5.2: Procedure to find a confidence interval for a population mean when σ is unknown

Assumptions:

1. Simple random sample
2. Normal population or large sample
3. σ unknown

Step 1: For a confidence level of $1 - \alpha$, use the Table in Appendix-*G* to find $t_{\alpha/2}$ with $df = n - 1$, where n is the sample size

Step 2: The confidence interval for μ is from

$$\bar{X} - t_{\alpha/2} \bullet \frac{S}{\sqrt{n}} \text{ to } \bar{X} + t_{\alpha/2} \bullet \frac{S}{\sqrt{n}}$$

where $t_{\alpha/2}$ is found in step 1 and $\bar{X}$ and S are computed from the sample data.

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E5.7

Find the following confidence intervals for μ_d assuming that the populations of paired differences are normally distributed.

(*a*) $n = 9$, $\bar{X} = 25$, $S = 13$, confidence level 99%.

(*b*) $n = 26$, $\bar{X} = 13$, $S = 5$, confidence level 95%.

(*c*) $n = 12$, $\bar{X} = 35$, $S = 12$, confidence level 90%.

Solution:

(*a*) The *t*-value for $\nu = n - 1 = 9 - 1 = 8$ d.o.f. and 0.005 area in the right tail is 3.355 (from the table in Appendix-*G*). The 99% confidence interval with $t_{0.005,8} = 3.355$ is

$$\bar{X} - t_{\alpha/2,\nu} S/\sqrt{n} \le \mu \le \bar{X} + t_{\alpha/2,\nu} S/\sqrt{n}$$

$$25 - 3.355(13)/\sqrt{9} \le \mu \le 25 + 3.355(13)\sqrt{9}$$

$$25 - 14.538 \le \mu \le 25 + 14.538$$

$$10.462 \le \mu \le 39.538$$

The complete statement of the confidence interval, with the associated probability is

$$P(10.462 \le \mu \le 39.538) = 0.99$$

(*b*) Using the data given, a 95% confidence interval with $t_{0.025,25} = 2.060$ from the table in Appendix-*G*, is given

$$13 - 2.060(5)/\sqrt{26} \le \mu \le 13 + 2.060(5)/\sqrt{26}$$

$$13 - 2.02 \le \mu \le 13 + 2.02$$

$$10.98 \le \mu \le 15.02$$

The complete statement of the confidence interval with the associated probability is

$P(10.98 \le \mu \le 15.02) = 0.95$

(*c*) Using the data given, a 90% confidence interval with $t_{0.05,11} = 1.796$ from the table in Appendix-*G*, is given by

$$35 - 1.796(12)/\sqrt{12} \le \mu \le 35 + 1.796(12)/\sqrt{12}$$

$$35 - 0.16 \le \mu \le 35 + 0.16$$

$$34.84 \le \mu \le 35.16$$

The complete statement of the confidence interval with the associated probability is

$P(34.84 \le \mu \le 35.16) = 0.90$

Example E5.8

Measurements on the percentage of enrichment of 12 fuel rods were reported as follows:

2.94	3	2.9	2.75	3	2.95
2.75	2.95	2.92	2.81	3.05	2.8

Find a 99% two-sided confidence interval on the mean percentage of enrichment of fuel rods. Can we state the mean percentage of enrichment is 2.95%? Why?

SOLUTION:

Here $n = 12$ and $\Sigma X = 34.82$

$$\bar{X} = \frac{\Sigma X}{n} = \frac{34.82}{12} = 2.9017$$

Also $\Sigma(X_i - \bar{X})^2 = 0.112567$

Hence $S = \sqrt{\dfrac{\Sigma(X_i - \bar{X})^2}{n-1}} = \sqrt{\dfrac{0.112567}{12-1}} = 0.10116$

The 99% two-sided confidence interval on the mean per cent enrichment is found from

$$\bar{X} - t_{0.005,11}\left(\frac{S}{\sqrt{n}}\right) \le \mu \le \bar{X} + t_{0.005,11}\left(\frac{S}{\sqrt{n}}\right)$$

For $\alpha = 0.01$ and $n = 12$, $t_{\alpha/2,\, n-1} = t_{0.005,11} = 3.106$ (from the table in Appendix-*G*).

Therefore $2.9017 - 3.106\left(\dfrac{0.10116}{\sqrt{12}}\right) \le \mu \le 2.9017 + 3.106\left(\dfrac{0.10116}{\sqrt{12}}\right)$

$$2.9017 - 0.896625$$

or $2.005 \le \mu \le 3.798325$

We can state that the mean percentage of enrichment 2.95 is included in the 99% two-sided confidence interval.

5.5 CONFIDENCE INTERVAL ON THE VARIANCE OF A NORMAL DISTRIBUTION

Here, we accept S^2 as providing a good estimate for σ^2. For setting a confidence interval for σ^2, we make the following assumptions:

1. The observations are independent.
2. The parent population has a normal distribution.
3. The sample is a simple random sample.

Suppose that X is normally distributed with unknown mean μ and unknown variance σ^2. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n, let S^2 be the sample variance. We note that the sampling distribution of

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

is chi-square with $n - 1$ degrees of freedom. This distribution is shown in Fig. 5.3.

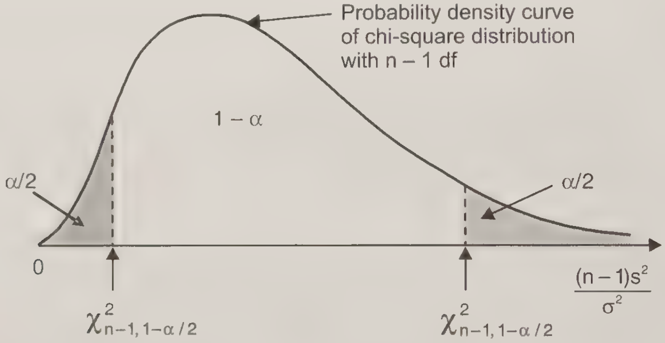


Fig. 5.3: Chi-square values such that areas 1 – α/2 and α/2 are to their right

We observe from Fig. 5.3 that,

$$P\left\{\chi^2_{1-\alpha/2,n-1} \le \chi^2 \le \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha$$

or
$$P\left\{\chi^2_{1-\alpha/2,n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha \tag{5.13}$$

Equation (5.13) can be rearranged to give

$$P\left\{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right\} = 1-\alpha \tag{5.14}$$

Comparing Eqs. (5.14) and (5.4), we see that a 100 (1 – α) % two-sided confidence interval on σ^2 is

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \tag{5.15}$$

when S^2 is the sample variance based on n observations from a normal population.

In order to find a 100 (1 – α)% lower-confidence interval on σ^2, set $U = \infty$ and replace $\chi^2_{\alpha/2,n-1}$ with $\chi^2_{\alpha,n-1}$, giving

$$\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}} \le \sigma^2 \tag{5.16}$$

The 100 (1 – α)% upper-confidence interval is found by setting $L = 0$ and replacing $\chi^2_{1-\alpha/2,n-1}$ with $\chi^2_{1-\alpha,n-1}$, resulting in

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}} \tag{5.17}$$

In all the above Eqs. (5.13 to 5.17), $\chi^2_{n-1,1-\alpha/2}$ and $\chi^2_{n-1,\alpha/2}$ represent values from the chi-square distribution with $n - 1$ degrees of freedom such that they leave, respectively, area of $1 - \alpha/2$ and $\alpha/2$ to their right, as shown in Fig. 5.3. The procedure to find a confidence interval for a population standard deviation is given in Table 5.3.

Table 5.3: Procedure to find a confidence interval for a population standard deviation, σ

Assumptions:

1. Simple random sample
2. Normal population

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*F* to find $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ with $df = n - 1$

Step 2: The confidence interval for σ is from

$$\sqrt{\frac{(n-1)}{\chi^2_{\alpha/2}}}\,S \quad \text{to} \quad \sqrt{\frac{(n-1)}{\chi^2_{1-\alpha/2}}}\,S$$

where $\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ are found in step 1, n is the sample size and S is computed from the sample data obtained

Step 3: Interpret the confidence interval.

Example E5.9

Data are collected on the driving time required to reach place *A* from place *B*. The driving time is assumed to follow a normal distribution. Twenty-one driving times are collected and the sample variance is calculated to be 2 hours. A 99% confidence interval on the true variance is desired.

SOLUTION:

The two-sided confidence interval is given by (Appendix-*F*)

$$\frac{(n-1)S^2}{\chi^2_{\alpha/2,\nu}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\nu}}$$

where $\nu = n - 1$.

From the table in Appendix-*F* with $\nu = 21 - 1 = 20$ and $\alpha/2 = 0.005$, we find that $\chi^2_{0.005,20} = 39.997 = \chi^2_{0.995,20} = 7.434$.

Therefore,

$$\frac{(21-1)(2^2)}{39.997} \le \sigma^2 \le \frac{(21-1)(2^2)}{7.434} \text{ or } 2 \le \sigma^2 \le 10.761$$

The complete statement of the confidence interval with the associated probability is

$$P(2 \text{ hrs} \le \sigma^2 \le 10.761 \text{ hrs}) = 0.99$$

Example E5.10

The capability of a specific gauge can be studied by measuring the weight of paper. The data for repeated measurements of one sheet of paper are given in Table E5.10. Construct a 95% one-sided, upper confidence interval for the standard deviation of these measurements.

Table E5.10

3.481	3.449	3.484	3.476	3.473
3.478	3.473	3.465	3.475	3.471
3.472	3.471	3.478	3.473	3.475

SOLUTION:

A 95% upper-confidence interval is found from Eq. (5.17) as follows:

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2_{0.95,14}}$$

$n = 15$, $S = 0.001354$ and $\chi^2_{0.95,14} = 6.57$ (from the table in Appendix-*F*).

$$\Sigma X_i = 52.094, \ \Sigma(X_i - \bar{X})^2 = 0.000901$$

$$S = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} = 0.001354 \text{ and } S^2 = 1.83 \times 10^{-6}$$

$$\sigma^2 = \frac{14(1.83 \times 10^{-6})}{6.57}$$

$$\sigma^2 \le 3.9073 \times 10^{-6}$$

The above can be converted into a confidence interval on the standard deviation σ by taking the square root of both sides, resulting in $\sigma = 0.00197667$.

Therefore, at the 95% level of confidence, the data indicates that the process standard deviation could be as large as 0.00197667.

5.6 CONFIDENCE INTERVAL ON A POPULATION PROPORTION

We will now give an interval estimate for population proportion under the following assumptions:

1. The sample consists of n independent observations.
2. The sample size is large.
3. The population proportion is not too close to 0 or 1.

If a random sample size n has been taken from a large (possibly infinite) population, and $X(\leq n)$ observations in this sample belong to a class of interest, then $\hat{p} = X/n$ is the point estimator of the proportion of the population belonging. n and p are the parameters of binomial distribution. The sampling distribution $\hat{p}$ is approximately normal mean p and variance $p(1-p)/n$, if p is not too close to either 0 or 1, and if n is relatively large. Therefore, the distribution of

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

is approximately standard normal distribution.

We also observe that

$$P\{-Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}\} = 1 - \alpha$$

or

$$P\left\{-Z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq Z_{\alpha/2}\right\} = 1 - \alpha \tag{5.18}$$

Rearranging Eq. (5.18)

$$P\left\{\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right\} = 1 - \alpha \tag{5.19}$$

The quantity of $\sqrt{p(1-p)/n}$ is called the *standard error of the point estimator* $\hat{p}$. Replacing p by $\hat{p}$ in the standard error, giving an estimated standard error.

$$P\left\{\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right\} = 1 - \alpha \tag{5.20}$$

The above formula, Eq. (5.20), is applicable only if the observed sample proportion is not too close to 0 or 1 and the observed number of successes X and the observed number of failures $(n - X)$ both exceed 5. If these conditions are not met, the procedure is not recommended.

The above procedure depends on the adequacy of the normal approximation to the binomial. Conservatively speaking, this requires that np and $n(1 - p)$ be greater than or equal to 5. In situations where the approximation is inappropriate (especially when n is small) other methods must be used. Tables of binomial distribution could be used to obtain a confidence interval for p.

We may find approximate one-sided confidence bounds on p by a simple modification of Eq. (5.20).

The approximate $100(1 - \alpha)\%$ lower and upper confidence bounds are

$$\hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \quad \text{and} \quad p \leq \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

respectively.

The procedure to find a confidence interval for a population proportion, p is given in Table 5.4.

Table 5.4: Procedure to find a confidence interval for a population standard proportion, p

Assumptions:

1. Simple random sample
2. The number of successes, x, and the number of failures, $n - x$ are both 5 or greater

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*E* to find $z_{\alpha/2}$

Step 2: The confidence interval for p is from

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ to } \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $z_{\alpha/2}$ is found in step 1, n is the sample size and $\hat{p} = x/n$ is the sample proportion

Step 3: Interpret the confidence interval.

Example E5.11

In a random sample of 300 automobile crankshaft bearings, 12 have a surface finish that is rougher than the specifications allow.

(*a*) Calculate a 95% two-sided confidence interval on the fraction of defective crankshafts produced by this particular process.

(*b*) Calculate a 95% upper confidence bound on the fraction of defective crankshafts.

SOLUTION:

(*a*) 95% confidence interval on the fraction defectives produced with the process.

$$\hat{p} = \frac{12}{300} = 0.04,\ n = 30,\ z_{\alpha/2} = 1.96 \text{ (from the table in Appendix-}E\text{).}$$

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.04 - 1.96\sqrt{\frac{0.04(1-0.04)}{300}} \le p \le 0.04 + 1.96\sqrt{\frac{0.04(1-0.04)}{300}}$$

$$0.01 - 0.022175 \le p \le 0.04 + 0.02275$$

$$0.017825 \le p \le 0.062175$$

(*b*) 95% upper confidence bound:

$$z_{\alpha/2} = z_{0.05} = 1.65 \text{ (from Appendix-}E\text{)}$$

$$p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$p \le 0.04 + 1.65\sqrt{\frac{0.04(1-0.04)}{300}}$$

or $\le 0.04 + 0.018668$

Hence $p \le 0.058668$

Example E5.12

A sample of 300 of electronic components produced showed 36 non-conforming ones. Construct a 95% confidence interval for this. Assume binomial distribution.

SOLUTION:

A two-sided confidence interval on p is given by

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Since n is large and the underlying distribution is binomial, $\hat{p} = 36/300 = 0.12$

A 95% confidence interval, with $Z_{0.025} = 1.96$ (from the table Appendix-*E*) is given by

$$0.12 - 1.96\sqrt{\frac{(0.12)(0.88)}{300}} \le p \le 0.12 + 1.96\sqrt{\frac{(0.12)(0.88)}{300}}$$

or

$$0.12 - 1.96\ (0.01876) \le p \le 0.12 + 1.96\ (0.01876)$$

$$0.12 - 0.0368 \le p \le 0.12 + 0.0368$$

$$0.0832 \le p \le 0.1568$$

The complete statement of the confidence interval with the associated probability is

$$P(0.0832 \le p \le 0.1568) = 0.95$$

Margin of Error:

The margin of error from Eq. (5.20) is given by

$$\text{margin of error} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

It can be shown that the maximum value of $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is $z_{\alpha/2}\sqrt{\frac{1}{4n}}$ and occurs when $\hat{p} = \frac{1}{2}$. This called the *safe* or *conservative margin of error*, whatever the actual value of $\hat{p}$ in the population.

Hence, conservative margin of error = $z_{\alpha/2}\sqrt{\frac{1}{4n}}$.

Example E5.13

Determine the margin of error in Example E5.12.

SOLUTION:

The margin of error is given by

$$\text{margin of error} = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*), $\hat{p} = 36/300 = 0.12$

Hence, margin of error = $1.96\sqrt{\dfrac{0.12(1-0.12)}{300}} = 0.0368$

5.7 CONFIDENCE INTERVAL ON THE DIFFERENCE IN TWO MEANS, VARIANCE KNOWN

The procedure presented in this section for finding a confidence interval is valid if the following assumptions are justified:

1. The two populations are normally distributed. This assumption is not very important if both sample sizes are at least 30 (large sample size).
2. The standard deviations of the two populations are known.
3. Two random samples are picked, one from each population. They are independent, that is, any outcome in one sample does not influence any outcome in the other sample.
4. Within each sample the outcomes are independent.

Consider two independent random variables X_1 with unknown mean μ_1 and known variance σ_1^2 and X_2 with unknown mean μ_2 and variance σ_2^2. We wish to find a 100 (1 – α)% confidence interval on the difference in means μ_1, μ_2. Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample of n_1 observations from X_1; and $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample of n_2 observations from X_2. If $\bar{X}_1$ and $\bar{X}_2$ are the sample means, the statistic

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

is standard normal if X_1 and X_2 are normal or approximately standard normal if the conditions of the central limit theorem apply, respectively. Also, σ_1 and σ_2 are the population standard deviations. From Fig. 5.2 it is clear that

$$P\{-z_{\alpha/2} \le z \le z_{\alpha/2}\} = 1 - \alpha$$

or
$$P\left\{-z_{\alpha/2} \le \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \le z_{\alpha/2}\right\} = 1 - \alpha$$

where $z_{\alpha/2}$ is the upper α/2 percentage point of the standard normal distribution.

Rearranging, we have

$$P\left\{\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right\} = 1 - \alpha \qquad (5.21)$$

Comparing Eqs. (5.4) and (5.21), we note that 100(1 – α)% confidence interval for $\mu_1 - \mu_2$ is

$$\bar{X}_1 - \bar{X}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (5.22)$$

One-sided confidence intervals on $\mu_1 - \mu_2$ may also be obtained. A 100 $(1 - \alpha)\%$ upper-confidence interval on $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (5.23)$$

and a $100(1 - \alpha)$ % lower-confidence interval is

$$\bar{X}_1 - \bar{X}_2 - z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \qquad (5.24)$$

The confidence level $(1 - \alpha)$ is exact when the populations are normal. For non-normal populations, the confidence level is approximately valid for large sample sizes.

The margin of error E is

$$E = z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (5.24a)$$

Example E5.14

The mean hourly wage for male workers was \$15 and \$13 for female workers in a particular manufacturing sector in the year 2000. These two estimations were based on random samples of 1000 and 1200 workers taken, respectively from two independent populations. The standard deviations of the two populations are known to be \$2 and \$1.50 respectively. Construct a 95% confidence interval for the difference between the mean hourly wages of the two populations.

SOLUTION:

For the 95% confidence level $z_{\alpha/2} = 1.96$ (from the table in Appendix-*E*).

The 95% confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$15 - 13 \pm 1.96\sqrt{\frac{2^2}{1000} + \frac{1.5^2}{1200}} = 2 \pm 1.96\,(0.2046) = 2 \pm 0.4011 = \$1.60 \text{ to } \$2.40$$

We have obtained (\$1.60, \$2.40) as an 95% confidence interval for $\mu_1 - \mu_2$.

Example E5.15

Two different formulations of oxygenated motor fuel are being tested to study their road octane numbers. The variances of road octane number for formulation *A* is $\sigma_1^2 = 1.5$ and for formulation *B* it is $\sigma_2^2 = 1.2$. Two random samples of sizes $n_1 = 15$ and $n_2 = 20$ are tested, and the mean road octane numbers observed are

$\bar{X}_1 = 90$ and $\bar{X}_2 = 93$. Assume normality and calculate a 95% confidence interval on the difference in means.

SOLUTION:

95% confidence interval: $\alpha = 0.05$, $z_{\alpha/2} = z_{0.025} = 1.96$ (from the table in Appendix-*E*).

$$\left(\bar{X}_1 - \bar{X}_2\right) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{X}_1 - \bar{X}_2\right) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$90 - 93 - 1.96\sqrt{\frac{1.5}{15}+\frac{1.2}{20}} \le \mu_1 - \mu_2 \le 90 - 93 + 1.96\sqrt{\frac{1.5}{15}+\frac{1.2}{20}}$$

$$-3 - 0.784 \le \mu_1 - \mu_2 \le -3 + 0.784$$

or $$-3.784 \le \mu_1 - \mu_2 \le -2.216$$

With 95% confidence, we believe the mean road octane number for formulation *B* exceeds that of formulation *A* since 0 is not included in the confidence interval.

5.8 CONFIDENCE INTERVAL ON THE DIFFERENCE IN MEANS OF TWO NORMAL DISTRIBUTIONS, VARIANCES UNKNOWN

We make the following assumptions:

1. The two populations are normally distributed.
2. The population variances σ_1^2 and σ_2^2, though unknown are the same, say, each equal to σ^2.
3. Two random and independent samples are picked, one from each population.

Consider two independent normal random variables, say X_1 with mean μ_1 and variance σ_1^2, and X_2 with mean μ_2 and variance σ_2^2. Both the means μ_1 and μ_2 and the variances σ_1^2 and σ_2^2 are unknown. Assume that both variances are equal; that is $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We wish to find a 100 (1 – α)% confidence interval on the difference in means $\mu_1 - \mu_2$.

Random samples of size n_1 and n_2 are taken on X_1 and X_2, respectively; let the sample means be denoted by $\bar{X}_1$ and $\bar{X}_2$, and the sample variances be denoted S_1^2 and S_2^2. Since both S_1^2 and S_2^2 are estimates of the common variance σ^2, we may obtain a combined (or "pooled") estimator of σ^2 as the pooled estimate of the common population standard deviation, S_p, is given by

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \tag{5.25}$$

To find the confidence interval for $\mu_1 - \mu_2$, note that the distribution of the statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

is the t-distribution with $n_1 + n_2 - 2$ degrees of freedom. Hence,

$$P\left\{-t_{\alpha/2,n_1+n_2-2} \leq t \leq t_{\alpha/2,n_1+n_2-2}\right\} = 1-\alpha$$

or

$$P\left\{-t_{\alpha/2,n_1+n_2-2} \leq \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \leq t_{\alpha/2,n_1+n_2-2}\right\} = 1-\alpha$$

Here $t_{\alpha/2,n_1+n_2-2}$ is the upper $\alpha/2$ percentage point of the t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Rearranging,

$$P\left\{\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right\} = 1-\alpha \qquad (5.26)$$

Hence, a 100 $(1 - \alpha)$% two-sided confidence interval on the difference in means $\mu_1 - \mu_2$ is

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \qquad (5.27)$$

A one-sided 100 $(1 - \alpha)$% lower-confidence interval on $\mu_1 - \mu_2$ is

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \leq \mu_1 - \mu_2 \qquad (5.28)$$

and a one-sided 100 $(1 - \alpha)$% upper-confidence interval on $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \leq \bar{X}_1 - \bar{X}_2 + t_{\alpha,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \qquad (5.29)$$

The margin of error E is given by

$$E = z_{\alpha/2}\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Since σ is not known, we replace it with its pooled estimate S_p and thereby we cannot use z-values from the normal table. Here, we use values from the t-distribution with $n_1 + n_2 - 2$ degrees of freedom. This is especially important if the degrees of freedom are less than 30 because then the t-values will be markedly different from the z-values. Hence, the marginal of error is

$$E = t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

where $$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{\Sigma(X_{1i} - \bar{X}_1)^2 + \Sigma(X_{2i} - \bar{X}_2)^2}{n_1+n_2-2}}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the sample means, based, respectively on n_1 and n_2 observations, and S_p is the pooled estimate of σ. The procedure to find a confidence interval for the difference between two population means, μ_1 and μ_2 is given in Table 5.5.

Table 5.5: Procedure to find a confidence interval for the difference between two population means μ_1 and μ_2

Assumptions:

1. Simple random samples
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*G* to find $t_{\alpha/2}$ with $df = n_1 + n_2 - 2$

Step 2: The end points of the confidence interval for $\mu_1 - \mu_2$ are

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{where } S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E5.16

The overall distance travelled by a golf ball is tested by striking the ball with a specially made mechanical golfer. Ten randomly selected balls of two different brands *A* and *B* are tested and the overall distance measured as shown below:

Brand *A*	263	267	271	273	275	276	279	283	286	287
Brand *B*	244	258	261	263	265	268	270	271	273	280

Construct a 95% two-sided confidence interval on the mean difference in overall distance between the two brands of golf balls.

SOLUTION:

Here $n_1 = 10, n_2 = 10, \Sigma X_{1i} = 2760, \Sigma X_{2i} = 2653$

$$\bar{X}_1 = \frac{\Sigma X_{1i}}{n_1} = \frac{2760}{10} = 276, \quad \bar{X}_2 = \frac{\Sigma X_{2i}}{n_2} = \frac{2653}{10} = 265.3$$

$$\Sigma(X_{1i} - \bar{X}_1)^2 = 564 \text{ and } \Sigma(X_{2i} - \bar{X}_2)^2 = 868.1$$

$$S_1^2 = \frac{\Sigma(X_{1i} - \bar{X}_1)^2}{n_1 - 1} = 62.6667 \text{ and } S_1 = 7.9162$$

$$S_2^2 = \frac{\Sigma(X_{2i} - \bar{X}_2)^2}{n_2 - 1} = 96.4556 \text{ and } S_2 = 9.8212$$

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} = \sqrt{\frac{(10-1)(62.6667) + (10-1)(96.4556)}{10 + 10 - 2}}$$

or $\quad S_p = 8.9197$

95% two-sided confidence interval: $t_{\alpha/2,n_1+n_2-2} = t_{0.025.18} = 2.101$ (from the table in Appendix-*G*).

$$\left(\bar{X}_1 - \bar{X}_2\right) - t_{\alpha/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \le \mu_1 - \mu_2 \le \left(\bar{X}_1 - \bar{X}_2\right) + t_{\alpha/2,n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$(276 - 265.3) - 2.101(8.9197)\sqrt{\frac{1}{10} + \frac{1}{10}} \le \mu_1 - \mu_2 \le (276 - 265.3) + 2.101(8.9197)\sqrt{\frac{1}{10} + \frac{1}{10}}$$

$$2.3191 \le \mu_1 - \mu_2 \le 19.0809.$$

Example E5.17

To compare the tensile strength of two materials, 12 test specimens of each material were tested. The order was decided at random. The higher the test value, the better the strength of these materials. The data obtained are summarised in Table E5.20.

Table E5.20

Material A	12.5	11.7	9.70	9.60	10.5	9.60	9.40	11.5	8.70	11.5	10.6	9.7
Material B	9.40	8.40	11.6	7.20	9.70	7.00	10.4	8.20	7.30	9.20	6.90	12.7

SOLUTION:

For the both materials

Material *A*	Material *B*
$n = 12$	$n = 12$
$\Sigma X_{1i} = 124.8$	$\Sigma X_{2i} = 108$
$\bar{X}_1 = \frac{\Sigma X_{1i}}{n} = \frac{124.8}{12} = 10.4$	$\bar{X}_2 = \frac{\Sigma X_{2i}}{n} = \frac{108}{12} = 9$
$\Sigma X_{1i}^2 = 1312$	$\Sigma X_{2i}^2 = 1010.64$
$\left(\frac{1}{n_1}\right)(\Sigma X_{1i})^2 = 1297.92$	$\left(\frac{1}{n_2}\right)(\Sigma X_{2i})^2 = 972$
$\Sigma\left(X_{1i} - \bar{X}_1\right)^2 = 14.08$	$\Sigma\left(X_{2i} - \bar{X}_2\right)^2 = 38.64$

Hence, the summary table is

Material	n	Degree of freedom	$\bar{X}$	Sum of squares
A	12	11	10.4	14.08
B	12	11	9	38.64

Total: 52.72

Therefore, $d = \bar{X}_1 - \bar{X}_2 = 1.4$

The pooled estimate of variance is $S^2 = 52.72/22 = 2.4$

The estimated variance of the difference between the means of

$$S_d^2 = S^2\left(\frac{1}{n} + \frac{1}{n}\right) = 2.4\left(\frac{2}{12}\right) = 0.4$$

The estimated standard deviation of d is

$$S_d = \sqrt{0.4} = 0.63$$

Hence, if we desire a 95% confidence interval, we must obtain $t_{0.025,22} = 2.074$ (from the table in Appendix-*G*). The 95% confidence interval now be formed using the relationship

$$L = d - t_{\alpha/2, n_1+n_2-2}\ S_d = 1.40 - (2.074)(0.63) = 0.09$$

$$U = d + t_{\alpha/2, n_1+n_2-2}\ S_d = 1.40 + (2.074)(0.63) = 2.71$$

We can be a 95% confidence that the difference between the actual means is between 0.09 and 2.71.

When we cannot assume $\sigma_1^2 = \sigma_2^2$, we can find a 100(1 – α)% confidence interval on $\mu_1 - \mu_2$ using the fact that the statistic

$$t^* = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

distributed approximately as t with degrees of freedom given by

$$\nu = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(S_2^2/n_2\right)^2}{n_2 - 1}} \tag{5.30}$$

Therefore, an approximate 100 (1 – α)% two-sided confidence interval on $\mu_1 - \mu_2$ when $\sigma_1^2 \neq \sigma_2^2$ is

$$\bar{X}_1 - \bar{X}_2 - t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le \bar{X}_1 - \bar{X}_2 + t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{5.31}$$

Upper (lower) one-sided confidence limits may be found by replacing the lower (upper) confidence limit with $-\infty$ (∞) and changing $\alpha/2$ to α.

The procedure to find a confidence interval for the difference between two population means, μ_1 and μ_2 is given in Table 5.6.

Table 5.6: Procedure to find a confidence interval for the difference between two population means μ_1 and μ_2

Assumptions:

1. Simple random samples
2. Independent samples
3. Normal populations or large samples

Step 1: For a confidence level of $1-\alpha$, use the table in Appendix-*G* to find $t_{\alpha/2}$ with $df = \nu$, where

$$\nu = \frac{\left[\left(S_1^2/n_1\right)+\left(S_2^2/n_2\right)\right]^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1-1}+\dfrac{\left(S_2^2/n_2\right)^2}{n_2-1}}$$

rounded down to the nearest integer.

Step 2: The end points of the confidence interval for $\mu_1 - \mu_2$ are

$$\left(\bar{X}_1-\bar{X}_2\right)\pm t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

Step 3: Interpret the confidence interval.

Example E5.18

The administration of a college wants to test if the mean GPAs (grade point averages) of all male and female students who actively participate in sports are different. A random sample of 28 male students and 24 female students who are actively participated in sports. The mean GPAs of the two groups were found to be 2.00 and 2.12 respectively, with the corresponding standard deviations equal to 0.43 and 0.38. Assume the GPAs of all male and female students who participated in sports activities have a normal distribution with unequal standard deviations. Construct a 90% confidence interval for the difference between the two population means.

SOLUTION:

Given $S_1 = 0.43, S_2 = 0.38, n_1 = 28, n_2 = 24$

The desired confidence interval is given by

$$\left(\bar{X}_1-\bar{X}_2\right)\pm t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

where the degrees of freedom

$$\nu = \frac{\left[\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right]^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1-1}+\dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2-1}} = \frac{\left[\dfrac{(0.43)^2}{28}+\dfrac{(0.38)^2}{24}\right]^2}{\dfrac{\left(\dfrac{0.43^2}{28}\right)^2}{28-1}+\dfrac{\left(\dfrac{0.38^2}{24}\right)^2}{24-1}} = 49 \text{ (rounded)}$$

Area in each tail of the t-distribution curve = $0.5 - \frac{0.90}{2} = 0.05$. The t-value for $df = 49$ and 0.05 area in the right tail is 1.677 (from the table in Appendix-G).

The 90% confidence interval for $\mu_1 - \mu_2$ is given by

$$(2.0 - 2.12) \pm 1.677\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

or $$-0.12 \pm 1.677\sqrt{\frac{(0.43)^2}{28} + \frac{(0.38)^2}{24}}$$

$$-0.12 \pm 1.677(0.11234) = -0.12 \pm 0.19 = -0.31 \text{ to } 0.07.$$

Example E5.19

Assuming that two populations are normally distributed with unequal and unknown population standard deviations, construct a 99% confidence interval for $\mu_1 - \mu_2$ for the following data:

$$n_1 = 15,\ \bar{X}_1 = 50,\ S_1 = 3.0,\ n_2 = 19,\ \bar{X}_2 = 40,\ S_2 = 6.0$$

SOLUTION:

From the table in Appendix-G, $t_{0.005,22}$ is 2.756. The desired confidence interval is given by

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm t_{\alpha/2,\nu}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

where $$\nu = \frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Substituting the values for S_1, S_2, n_1 and n_2 and simplifying, we get n = 29. Thus, the desired confidence interval is given by:

$$50 - 40 - 2.756\sqrt{\frac{3^2}{15} + \frac{6^2}{19}} \le \mu_1 - \mu_2 \le 50 - 40 + 2.756\sqrt{\frac{3^2}{15} + \frac{6^2}{19}}$$

$$10 - 2.756\ (1.579) \le \mu_1 - \mu_2 \le 10 + 2.756\ (1.579)$$

$$10 - 4.353 \le \mu_1 - \mu_2 \le 10 + 4.353$$

$$5.647 \le \mu_1 - \mu_2 \le 14.353$$

The complete statement of the confidence level with the associated probability is

$$P(5.647 \le \mu \le 14.353) = 99$$

5.9 CONFIDENCE INTERVAL ON $\mu_1 - \mu_2$ FOR PAIRED OBSERVATIONS

In general, suppose that the data consists of n pairs (X_{11}, X_{21}), (X_{12}, X_{22}), …, (X_{1n}, X_{2n}). Both X_1 and X_2 are assumed to be normally distributed with mean μ_1 and μ_2, respectively. The random variables within different pairs are independent. Since, there are two measurements on the same experimental unit, the two measurements within the same pair may not be independent. Consider the n differences $D_1 = X_{11} - X_{21}$, $D_2 = X_{12} - X_{22}$, …, $D_n = X_{1n} - X_{2n}$. Now the mean of the differences D, say μ_D, is

$$\mu_D = E(D) = D(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2$$

because the expected value of $X_1 - X_2$ is the difference in expected values regardless of whether X_1 and X_2 are independent. Hence, we can construct a confidence interval on $\mu_1 - \mu_2$ by just finding a confidence interval on μ_D. To construct the confidence interval for $\mu_D = \mu_1 - \mu_2$, note that

$$T = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}}$$

follows a t-distribution with $n - 1$ degrees of freedom. Then, $P(-t_{\alpha/2,n-1} \le T \le t_{\alpha/2,n-1}) = 1 - \alpha$, we can substitute for T in the above expression and perform the necessary steps to isolate $\mu_D = \mu_1 - \mu_2$ between the inequalities. This leads to the following 100 $(1 - \alpha)$% confidence interval on $\mu_1 - \mu_2$. The differences D_i are normally and independently distributed, we can apply the t-distribution procedure described earlier to find the confidence interval on μ_D. By analogy with Eq. (5.10), the 100 $(1 - \alpha)$% confidence interval on $\mu_D = \mu_1 - \mu_2$ is

$$\bar{D} - t_{\alpha/2,n-1} s_D / \sqrt{n} \le \mu_D \le \bar{D} + t_{\alpha/2,n-1} s_D / \sqrt{n} \tag{5.32}$$

where $\bar{D}$ and s_D are the sample mean and sample standard deviation of the differences D_i, respectively. This confidence interval is valid for the case where $\sigma_1^2 \ne \sigma_2^2$, because s_D^2 estimates $\sigma_D^2 = V(X_1 - X_2)$ and for large samples ($n \ge 30$ pairs) and the assumption of normality is not necessary.

Summarising, if $\bar{d}$ and s_d are the sample mean and standard deviation of the difference of n random pairs of normally distributed measurements, a 100$(1 - \alpha)$% confidence interval on the difference in means $\mu_D = \mu_1 - \mu_2$ is

$$\bar{d} - t_{\alpha/2,n-1} s_d / \sqrt{n} \le \mu_D \le \bar{d} + t_{\alpha/2,n-1} s_d / \sqrt{n}$$

where $t_{\alpha/2,n-1}$ is the upper $\alpha/2$% of point of the t-distribution with $n - 1$ degrees of freedom.

This confidence interval is also valid for the case where $\sigma_1^2 \ne \sigma_2^2$, since $s_D^2 = V(X_1 - X_2)$. Also, for large samples (say, $n \ge 30$), the explicit assumption of normality is unnecessary because of the central limit theorem.

The procedure to find a confidence interval for the difference between two population means, μ_1 and μ_2 is given in Table 5.7.

Table 5.7: Procedure to find a confidence interval for the difference between two population means μ_1 and μ_2

Assumptions:

1. Simple random paired samples
2. Normal differences or large sample

Step 1: For a confidence level of $1 - \alpha$, use the table in Appendix-*G* to find $t_{\alpha/2}$ with $df = n - 1$,

Step 2: The end points of the confidence interval for $\mu_1 - \mu_2$ are

$$\bar{d} \pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$$

Step 3: Interpret the confidence interval.

The confidence interval is exact for normal differences and is approximately correct for large samples and normal differences.

Example E5.20

A company wanted to find out if attending a special course on "how to be a successful salesperson" can increase the average sales of its employees. The company sent 10 of its salespersons to attend this course. The following table gives the one-week sales of these salespersons before and after they attended this course.

Before	11	18	25	10	14	16	17	18	19	21
After	17	25	26	15	20	21	22	23	26	24

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the sales before and after attending the special course. Assume that the population of paired differences is approximately normally distributed.

SOLUTION:

$n = 10$

	Before	After	Difference, d	d^2
	11	17	−6	36
	18	25	−7	49
	25	26	−1	1
	10	15	−5	25
	14	20	−6	36
	16	21	−5	25
	17	22	−5	25
	18	23	−5	25
	19	26	−7	49
	21	24	−3	9
Σ	169	219	−50	280

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-50}{10} = -5$$

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{280 - \frac{(-50)^2}{10}}{10-1}} = 1.8257$$

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{1.8257}{\sqrt{10}} = 0.5774$$

$$df = n - 1 = 10 - 1 = 9;\ \alpha/2 = 0.025$$

From the table in Appendix-*G*, $t_{0.025,9} = 2.262$.

The 95% confidence interval for μ_d is

$$\bar{d} \pm t s_{\bar{d}} = -5 \pm 2.262(0.5774)$$

or -6.3060 to -3.6940.

Example E5.21

A college claims that the Math tutoring service offers significantly increases the test scores of students in mathematics. The following table gives the scores per 120 of 12 students before and after they took the tutoring help.

Before	81	75	88	90	67	71	92	69	73	74	75	79
After	96	74	93	79	79	73	111	76	75	78	79	84

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the score before attending the tutoring service minus the score after attending the tutoring service. Assume that the population of paired differences is approximately normally distributed.

SOLUTION:

$n = 12$

Before	After	Difference, d	d^2
81	96	−15	225
75	74	1	1
88	93	−5	25
90	79	11	121
67	79	−12	144
71	73	−2	4
92	111	−19	361
69	76	−7	49
73	75	−2	4
74	78	−4	16
75	79	−4	16
79	84	−5	25
934	997	−63	991

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-63}{12} = -5.25$$

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{991 - \frac{(-63)^2}{12}}{12-1}} = 7.7474$$

$$s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{7.7474}{\sqrt{12}} = 2.2365$$

$$df = n - 1 = 12 - 1 = 11;\ \alpha/2 = 0.05 - \frac{0.95}{2} = 0.025$$

From the table in Appendix-*G*, $t_{11.0.025} = 2.201$

The 95% confidence interval for μ_d is

$$\pm t = -5.25 + (2.201)(2.2365)$$

or $$-10.1725 \text{ to } -0.3275$$

5.10 CONFIDENCE INTERVAL ON THE RATIO OF VARIANCE OF TWO NORMAL DISTRIBUTIONS

Let X_1 and X_2 are independent normal random variables with unknown means μ_1 and μ_2 and unknown variances σ_1^2 and σ_2^2, respectively. We wish to find a 100 (1 – α)% confidence interval on the ratio of σ_1^2/σ_2^2. Let two random samples of sizes n_1 and n_2 are taken on X_1 and X_2, and let S_1^2 and S_2^2 denote the sample variances. The sampling distribution of

$$F = \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2}$$

is *F* with $n_2 - 1$ and $n_1 - 1$ degrees of freedom. This distribution is shown in Fig. 5.4.

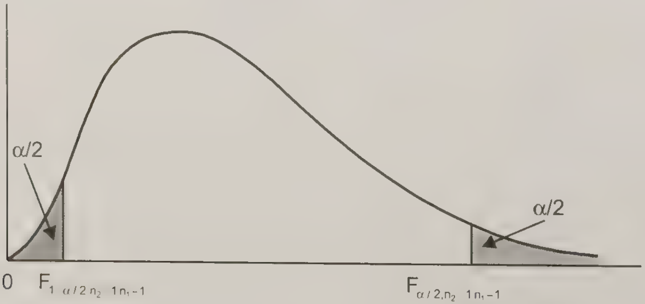


Fig. 5.4: The distribution of F_{n_2-1,n_1-1}

From Fig. 5.4, we observe that

$$P\left\{F_{1-\alpha/2,n_2-1,n_1-1} \le F \le F_{\alpha/2,n_2-1,n_1-1}\right\} = 1-\alpha$$

or
$$P\left\{F_{1-\alpha/2,n_2-1,n_1-1} \le \frac{S_2^2/\sigma_2^2}{S_1^2/\sigma_1^2} \le F_{\alpha/2,n_2-1,n_1-1}\right\} = 1-\alpha$$

Hence
$$P\left\{\frac{S_1^2}{S_2^2}F_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha/2,n_2-1,n_1-1}\right\} = 1-\alpha \tag{5.33}$$

Comparing Eqs. (5.33) and (5.4), we see that a 100 (1 – α)% two-sided confidence interval on σ_1^2/σ_2^2 is

$$\frac{S_1^2}{S_2^2}F_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha/2,n_2-1,n_1-1} \tag{5.34}$$

where the lower 1 – α/2 tail-point of the F_{n_2-1,n_1-1} distribution is given by

$$F_{1-\alpha/2,n_2-1,n_1-1} = \frac{1}{F_{\alpha/2,n_1-1,n_2-1}} \tag{5.35}$$

We can also construct one-sided confidence intervals. A 100 (1 – α)% lower-confidence limit on σ_1^2/σ_2^2 is

$$\frac{S_1^2}{S_2^2}F_{1-\alpha,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2} \tag{5.36}$$

while a 100(1 – α)% upper-confidence interval on σ_1^2/σ_2^2 is

$$\frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha,n_2-1,n_1-1} \tag{5.37}$$

where $F_{\alpha/2,n_2-1,n_1-1}$ and $F_{1-\alpha/2,n_2-1,n_1-1}$ are the upper and lower α/2 percentage points of the *F*-distribution with $n_2 - 1$ numerator and $n_1 - 1$ denominator degrees of freedom respectively. A confidence interval on the ratio of the standard deviations can be obtained by taking the square roots in Eq. (5.34).

Example E5.22

Two samples are drawn from two independent populations, which are normally distributed. A comparison of the samples is to be made by developing a 99% confidence interval on the ratio of their variances. The following information is known: $n_1 = n_2 = 21$, $S_1^2 = 5$ and $S_2^2 = 8$. Determine the complete statement of the confidence interval, with the associated probability.

SOLUTION:

A 100 (1 – α)% two-sided confidence interval on σ_1^2/σ_2^2 is given by

$$\frac{S_1^2}{S_2^2}F_{1-\alpha/2,\nu_2,\nu_1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{S_1^2}{S_2^2}F_{\alpha/2,\nu_2,\nu_1}$$

where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$

Therefore, $F_{0.025,20,20} = 2.46$ and $F_{0.975,19,19} = 1/2.46 = 0.4065$ (from the table in Appendix-*H*).

Confidence interval is given by

$$\frac{5}{8}(0.4065) \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{5}{8}(2.46)$$

or

$$0.254 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.538$$

The complete statement of the confidence interval with the associated probability is

$$P\left(0.254 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.538\right).$$

Example E5.23

The diameter of steel rods manufactured on two different machines is being investigated. Two random samples of sizes $n_1 = 15$ and $n_2 = 17$ are selected and the sample means and sample variations are: $\bar{X}_1 = 9$, $S_1^2 = 0.3$, $\bar{X}_2 = 8.95$ and $S_2^2 = 0.40$, respectively. Assume that the data is drawn from a normal distribution. Construct the following:

(a) a 90% two-sided confidence interval on σ_1/σ_2

(b) a 95% two-sided confidence interval on σ_1/σ_2. Comment on the comparison of the width of this interval with width of the interval in part *(a)*

(c) a 90% lower-confidence bound on σ_1/σ_2.

SOLUTION:

(a) $n_1 = 15, n_2 = 17$

$$f_{1-\alpha/2,\, n_1-1,\, n_2-1} = 0.412$$

$$f_{\alpha/2,\, n_1-1,\, n_2-1} = 2.33$$

90% confidence interval for the ratio of variances:

$$\left(\frac{S_1^2}{S_2^2}\right) f_{1-\alpha/2, n_1-1, n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \left(\frac{S_1^2}{S_2^2}\right) f_{\alpha/2, n_1-1, n_2-1}$$

$f_{0.05,14,17} = 2.33$ and $f_{0.95,14,16} = 0.412$ (from the table in Appendix-*H*)

$$\left(\frac{0.3}{0.4}\right) 0.412 \le \frac{\sigma_1^2}{\sigma_2^2} \le \left(\frac{0.3}{0.4}\right) 2.33$$

$$0.309 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.7475$$

$$0.55588 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.32193$$

(b) 95% confidence interval:

$$\left(\frac{S_1^2}{S_2^2}\right) f_{1-\alpha/2,n_1-1,n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2} \le \left(\frac{S_1^2}{S_2^2}\right) f_{\alpha/2,n_1-1,n_2-1}$$

$f_{0.025,14,16} = 2.82$ and $f_{0.975,14,16} = 0.342$ (from the table in Appendix-*H*)

$$\left(\frac{0.3}{0.4}\right) 0.342 \le \frac{\sigma_1^2}{\sigma_2^2} \le \left(\frac{0.3}{0.4}\right) 2.82$$

$$0.2565 \le \frac{\sigma_1^2}{\sigma_2^2} \le 2.115$$

$$0.50646 \le \frac{\sigma_1}{\sigma_2} \le 1.454304$$

The 95% confidence interval is wider than the 90% confidence interval.

(c) 90% lower-sided confidence interval:

$$\left(\frac{S_1^2}{S_2^2}\right) f_{1-\alpha/2,n_1-1,n_2-1} \le \frac{\sigma_1^2}{\sigma_2^2}$$

$f_{0.1,14,16} = 0.5$ (from the table in Appendix-*H*)

$$\left(\frac{0.3}{0.4}\right) 0.5 \le \frac{\sigma_1^2}{\sigma_2^2}$$

$$0.375 \le \frac{\sigma_1^2}{\sigma_2^2}$$

or

$$0.61237 \le \frac{\sigma_1}{\sigma_2}.$$

5.11 CONFIDENCE INTERVAL ON THE DIFFERENCE IN TWO PROPORTIONS

Here, we construct a confidence interval for $p_1 - p_2$ under the following assumptions:

1. The population proportions $\hat{p}_1$ and $\hat{p}_2$ are not too close to 0 or 1.
2. Two random samples are taken, one from each population, and the two samples are independent.
3. Sample sizes n_1 and n_2 are large.

If two independent samples of size n_1 and n_2 are taken from infinite populations so that X_1 and X_2 are independent, binomial random variables with parameters (n_1, p_1) and (n_2, p_2), respectively, where X_1 represents the number of sample observations from the first population that belongs to a class of interest and X_2 represents the number of sample observations from the second population that belongs to a class of interest, then $\hat{p}_1 = X_1/n_1$ and $\hat{p}_2 = X_2/n_2$ are independent estimator of p_1 and p_2 respectively. Considering the fact that the normal approximation to the binomial applies, the statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}}$$

is distributed approximately as standard normal. An approximate 100 (1 – α)% two-sided confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{5.38}$$

An approximate 100 (1 – α)% lower-confidence interval for $p_1 - p_2$ is

$$\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \le p_1 - p_2$$

and an approximate 100 (1 – α)% upper-confidence interval for $p_1 - p_2$ is

$$p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $z_{\alpha/2}$ is the upper α/2 percentage point of the standard normal distribution. The procedure to find a confidence interval for the difference between two population proportions, p_1 and p_2 is given in Table 5.8. The margin of error is

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Table 5.8: Procedure to find a confidence interval for the difference between two population proportions p_1 and p_2

Assumptions:

1. Simple random samples
2. Independent samples
3. x_1, $n_1 - X_1$, x_2 and $n_2 - X_2$ are all 5 or greater

Step 1: For a confidence level of 1 – α, use the table in Appendix-*E* to find $z_{\alpha/2}$

Step 2: The end points of the confidence interval for $p_1 - p_2$ are

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

where $z_{\alpha/2}$ is the upper α/2 percentage point of the standard normal distribution.

Step 3: Interpret the confidence interval.

Summary of confidence intervals for single population and two populations are given in Tables 5.9 and 5.10 respectively.

Example E5.24

An investigation on body mass index (bmi – a measure of body fat based on height and weight) for adults, a bmi of greater than 25 indicates an above health weight (i.e., overweight or obese). Of 800 randomly selected adults whose highest degree is bachelors, 412 have an above healthy weight; and of 1000 randomly selected adults with a graduate degree, 474 have an above healthy weight.

(a) determine a 90% confidence interval for the difference between the percentages of adults in the two degree categories who have an above healthy weight.

(b) repeat part (*a*) for an 80% confidence interval.

SOLUTION:

(a) Population 1: Bachelors degree, $\hat{p}_1 = \frac{412}{800} = 0.515$

Population 2: Graduate degree, $\hat{p}_2 = \frac{474}{1000} = 0.474$

$z_{\alpha/2} = 1.645$ (from the table in Appendix-*E*)

90% confidence interval is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$(0.515 - 0.474) \pm 1.645\sqrt{\frac{0.515(1-0.515)}{800} + \frac{0.474(1-0.474)}{1000}}$$

$$(0.515 - 0.474) \pm 1.645\sqrt{9.752813\times 10^{-4} + 2.49324\times 10^{-4}}$$

0.041 ± 0.057566 or -0.016566 to 0.098566

(b) The 80% confidence interval is given by

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\ ;\ z_{\alpha/2} = 1.28 \text{ (from Appendix-}E\text{)}$$

$$(0.515 - 0.474) \pm 1.28\sqrt{\frac{0.515(1-0.515)}{800} + \frac{0.474(1-0.474)}{1000}}$$

0.041 ± 0.04793 or 0.0037929 to 0.08893.

Table 5.9: Summary of confidential intervals (single population)

No.	Nature of the population	Parameter on which confidence intervals is set	Procedure	Limits of confidence interval with confidence coefficient $1-\alpha$
1.	Quantitative data; standard deviation σ is known; Population normal	μ, the population mean	Draw a sample of size n and compute the value of $\bar{X}$, the estimate of μ	$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$

2.	Quantitative data; standard deviation σ is not known; population normal; important when sample size is small ($n < 30$)	μ, the population mean	Draw a sample of size n and compute $\bar{X}$ and $S = \sqrt{\frac{1}{n-1}\Sigma(X_i - \bar{X})^2}$	$\bar{X} \pm t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}$ $t_{n-1,\,\alpha/2}$ is the value obtained from the t-distribution with $n - 1$ degrees of freedom
3.	Quantitative data; standard deviation σ is not known; population not necessarily normal; sample size is large ($n \geq 30$)	μ, the population mean	Draw a sample of size n and compute $\bar{X}$ and S	$\bar{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}$ Confidence interval is approximate
4.	Quantitative data; binomial case	p, the probability of success (the population proportion)	Draw a sample of size n and note X, the number of successes; obtain X/n, the estimate of p. Here n is assumed large	$\frac{X}{n} \pm Z_{\alpha/2}\sqrt{\frac{\frac{x}{n}\left(1-\frac{X}{n}\right)}{n}}$ Confidence interval is based on the central limit theorem, thus, is approximate
5.	Quantitative data; population normal	σ^2, the population variance	Draw a sample of size n and compute s^2	$\left(\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right)$

Table 6.10: Summary of confidence intervals (two populations)

No.	Nature of the population	Parameter on which confidence intervals is set	Procedure	Limits of confidence interval with confidence coefficient $1 - \alpha$
1.	Quantitative data; variances σ_1^2 and σ_2^2 are known; both populations are normally distributed	$\mu_1 - \mu_2$, the difference of the population means	Draw samples of sizes m and n from the two populations, get the respective means $\bar{X}_1$ and $\bar{X}_2$ and find the value of $\bar{X}_1 - \bar{X}_2$, an estimate of $\mu_1 - \mu_2$	$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{m} + \frac{\sigma_2^2}{n}}$
2.	Quantitative data; σ_1^2, σ_2^2 are not known but assumed equal; both populations are normally distributed	$\mu_1 - \mu_2$, the difference of the population means	Draw samples of sizes n_1 and n_2 from the two populations, compute $\bar{X}_1 - \bar{X}_2$ and $S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{m+n-2}}$	$(\bar{X}_1 - \bar{X}_2) \pm t_{m+n-2,\alpha/2} S_p \sqrt{\frac{1}{m} + \frac{1}{n}}$

No.	Nature of the population	Parameter on which confidence intervals is set	Procedure	Limits of confidence interval with confidence coefficient $1-\alpha$
3.	Quantitative data; σ_1^2 and σ_2^2 are not known and not assumed equal; populations may not be normal; sample sizes m and n are large	$\mu_1 - \mu_2$, the difference of the population means	Draw samples of sizes n_1 and n_2 compute $\bar{X}_1 - \bar{X}_2$; compute S_1^2 and S_2^2	$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$ Confidence interval is approximate
4.	Quantitative data μ_1 and μ_2 are the means of populations 1 and 2 respectively. Standard deviation of differences are unknown	$\mu_D = \mu_1 - \mu_2$ Difference in means of two normal distributions in a paired analysis	Paired analysis calculates differences and uses one sample method for inference on the mean difference	$\bar{d} - t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}} \le \mu_D$ $\le \bar{d} + t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}}$
5.	Quantitative data; σ_1^2 and σ_2^2 are the variances of population 1 and 2; σ_1^2 and σ_2^2 known	σ_1^2/σ_2^2 Ratio of variances of two normal distribution	Draw 2 samples, one from each population, compute S_1^2 and S_2^2	$\frac{S_1^2}{S_2^2} f_{1-\alpha/2,n_2-1,n_1-1} \le \frac{\sigma_1^2}{\sigma_2^2}$ $\le \frac{S_1^2}{S_2^2} f_{\alpha/2,n_2-1,n_1-1}$
6.	Quantitative data; p_1 is the proportion in population 1 and p_2 in population 2; sample sizes n_1 and n_2 are assumed large	$p_1 - p_2$, the difference of the population proportions	Draw two samples, one from each population, of sizes n_1 and n_2; find $\frac{\chi_1}{n_1}$ and $\frac{\chi_2}{n_2}$, the estimates of p_1 and p_2 respectively; then get $\frac{\chi_1}{n_1} - \frac{\chi_2}{n_2}$	$\frac{\chi_1}{n_1} - \frac{\chi_2}{n_2} \pm$ $Z_{\alpha/2}\sqrt{\frac{\frac{\chi_1}{n_1}\left(1-\frac{\chi_1}{n_1}\right)}{n_1}+\frac{\frac{\chi_2}{n_2}\left(1-\frac{\chi_2}{n_2}\right)}{n_2}}$ Confidence interval is based on the central limit theorem, thus is approximate

5.12 SAMPLE SIZE SELECTION

5.12.1 Sample Size Selection for Estimating Population Mean

Each level is constructed with regard to a given *confidence level* and is called a *confidence interval*. The confidence level associated with a confidence interval states how much confidence we have that this interval contains the true population parameter. The confidence level is denoted by 100 $(1 - \alpha)\%$. When expressed as probability, it is called the *confidence coefficient* and is denoted by $(1 - \alpha)$. α is called the *significance level*.

The confidence interval for μ for large samples is given as follows:

The 100 $(1 - \alpha)\%$ confidence interval for μ is

$$\bar{x} \pm z\sigma_x \qquad \text{if } \sigma \text{ is known}$$

$$\bar{x} \pm zs_x \qquad \text{if } \sigma \text{ is unknown} \tag{5.39}$$

where $\sigma_x = \sigma\sqrt{n}$ and $s_x = s/\sqrt{n}$

The value of z used here is obtained from standard normal distribution table for the given confidence level. The quantity $z\sigma_x$ (or zs_x when σ is not known) in the confidence interval formula is called the *maximum error of estimate* and is denoted by E.

Thus, $E = zs_x$ or $E = z\sigma_x$

The standard deviation of the sample mean is $\sigma/\sqrt{n}$

$$E = z\frac{\sigma}{\sqrt{n}} \tag{5.40}$$

Given the confidence level and the standard deviation of the population, the sample size that will produce a predetermined error E of the confidence interval estimate of μ is

$$n = \frac{z^2\sigma^2}{E^2} \tag{5.41}$$

where n = required sample size.

We note that

1. As the desired length of the interval $2E$ decreases, the required sample size n increases for a fixed value of σ and specified confidence.
2. As σ increases, the required sample size n increases for a fixed desired length $2E$ and specified confidence.
3. As the level of confidence increases, the required sample size n increases for fixed desired length $2E$ and standard deviation σ.

Example E5.25

At a 99% confidence level, the standard deviation σ for the population is 0.7. How large a sample should be selected if we want the estimate to be within 0.01 of the population mean?

SOLUTION:

The mean size is $\bar{x} \pm 0.01$

Therefore, the maximum size of the error of the estimate is to be 0.01, that is $E = 0.01$. The value of z for a 99% confidence level is 2.58. The value of σ is given to be 0.7. Hence, substituting the values and simplifying Eq. (5.41), we get

$$n = \frac{z^2\sigma^2}{E^2} = \frac{(2.58)^2(0.7)^2}{(0.01)^2} = 32,616$$

Hence, the required sample size is 32, 616.

5.12.2 Sample Size for the Estimation of Proportion

The maximum error, E, of the interval estimation of the population proportion is

$$E = z\sigma_{\hat{p}} = z\sqrt{\frac{pq}{n}}$$

The above expression can be manipulated algebraically to write it in terms of E, p, q and z and the final result is given here. Given the confidence level and the values of p and q, the sample size that will produce a predetermined maximum error E of the confidence interval estimate p is given by

$$n = \frac{z^2 pq}{E^2}$$

Another approach to choosing n uses the fact that the sample size from the above equation will always be a maximum for $p = 0.5$ ($q = 0.5$) or $p(1-p) \le 0.25$ with equality for $p = 0.5$. This can be used to obtain an upper bound on n. We are at least 100 $(1-\alpha)$% confident that the error in estimating p by $\hat{p}$ is less than E if the sample size is

$$n = \frac{Z^2}{E^2}(0.25)$$

Example E5.26

(*a*) How large a sample should be selected so that the maximum error of estimated for a 99% confidence interval for p is 0.04 when the value of the sample proportion obtained from a preliminary sample is 0.2?

(*b*) Find the most conservative sample size that will produce the maximum error for a 99% confidence interval for p equal to 0.04.

SOLUTION:

(*a*) Here $E = 0.04,\ \hat{p} = 0.2,\ \hat{q} = 1 - 0.2 = 0.8$

$z = 2.58$ for 99% confidence level (from the table in Appendix-E)

Hence $n = \dfrac{z^2\hat{p}\hat{q}}{E^2} = \dfrac{2.58^2(0.2)(0.8)}{(0.04)^2} = 66.564 \approx 67$

(*b*) $E = 0.04$, $z = 2.58$, and $p = q = 0.50$ for most conservative sample size.

$$\text{Thus, } n = \frac{z^2 pq}{E^2} = \frac{2.58^2(0.5)(0.5)}{(0.04)^2} = 1040.062 \approx 1040$$

5.13 SUMMARY

In this chapter, we introduced that part of statistics called *inferential statistics.* In Chapter 1, inferential statistics was described as the part of statistics that help us to make decisions about some characteristics of a population based on sample information. Inferential statistics makes use of the sample results to make decisions and draw conclusions about the population from which the sample is drawn. *Estimation* is the first topic to be covered in our presentation of inferential statistics. *Estimation* and *hypothesis testing* (presented in Chapter 6) topics taken together are usually referred to as *making inferences.* This chapter examined how to estimate the population mean, population proportion for a single population. We have extended our discussion of estimation to the difference between two population means and the difference between two population proportions.

PROBLEMS

P5.1 A manufacturing company tested 25 printed circuit boards and the findings are as follows: Mean length of life for 25 boards = 3,566 hours. Standard deviation of life length in sample = 150 hours. Construct a 90% confidence interval for the mean length of life for the new printed circuit boards.

P5.2 The mean tensile strength of a sample of 25 high quality steel specimens equals to 50,000 MPa. Find a 95% confidence interval on the mean tensile strength if the standard deviation is known to be 500 MPa.

P5.3 (*a*) The standard deviation of a normally distribution manufacturing process is 3g. Determine the sample size for a 90% confidence interval so that the estimation of the mean process is within 1g of the true but unknown mean yield.

(*b*) An electrical light bulb manufacturing company likes to estimate the light bulb's mean life. Assume a normal distribution of the life of a light bulb. Assuming the standard deviation is 30 hours, find the how many bulbs should be tested so as to be: *(i)* 95% confident that the estimate $\bar{x}$ will not differ from the true mean life by more than 10 hours. *(ii)* 98% confident to accomplish *(i)*.

P5.4 A college administrator wishes to estimate the mean age of all students currently enrolled. In a random sample of 20 students, the mean age is found to be 21 years. From past studies, the standard deviation is known to be 1.5 years. Construct a 90% confidence interval of the population mean age.

P5.5 In order to estimate the amount of time (in minutes) that a teller spends on a customer, a bank manager decided to observe 64 customers picked at random. The amount of time the teller spent on each customer was recorded. It was found that a sample mean was 4 minutes will standard deviation 1.2 minutes. Find a 98% confidence interval for the mean amount of time μ.

P5.6 When 25 cigarettes of a particular brand were tested in a laboratory for the amount of tar content, it was found that their mean content was 20 milligrams with $S = 2$ milligrams. Set a 90% confidence interval for the mean tar content μ in the population of cigarettes of the brand. Assume that the amount of tar in a cigarette is normally distributed.

P5.7 The breaking strength of a machine part is required to be at least 75 MPa. Based on previous test data, the standard deviation of breaking strength is 6 MPa. A random sample of 16 specimens is tested and the average value is found to be 73 MPa. Construct a 95% confidence interval on the mean breaking strength.

P5.8 The diameters of copper shafts produced by a certain manufacturing process should have a mean diameter of 0.51 mm. The diameter is known to have a standard deviation of 0.0002 mm. A random sample 40 shafts has an average diameter of 0.509 mm. Construct a 95% confidence interval on the mean shaft diameter.

P5.9 A normally distributed random variable has a known variance of $\sigma^2 = 36$ and unknown mean. Construct a 95% confidence interval on the mean that has a total width of 2.0.

P5.10 In an experimental lab, 10 students are selected at random from a class and were assigned to set up an experiment. Their time to complete the experimental set up are recorded (in minutes) as 51, 55, 54, 50, 56, 59, 57, 59, 58 and 60. Use a 95% confidence interval to estimate μ, the mean time for the student to set up the experiment. Assume that the population is normally distributed.

P5.11 A study claims that students in a particular college spend an average of 18 hours per week on leisure activities. The administration wanted to test this claim. A sample of 10 students selected at random gave their responses as follows (in hours):

14, 16, 19, 22, 23, 25, 26, 33, 38 and 41

Assume that the time spent on leisure activities by all the students is normally distributed. Use a 95% confidence interval activities.

P5.12 The time required to change flat tires in a garage is a normal distributed random variable. The tire changing times for 15 automobiles selected at random are as follows:

5, 9, 7, 6, 24, 11, 4, 13, 10, 9, 20, 8, 19, 17 and 25

Use a 95% confidence interval to estimate μ, the mean time required to change flat tires in that garage.

P5.13 A sample of 26 observations selected from a normally distributed population produced a sample variance of 35. Construct a confidence interval for σ^2 for each of the following confidence levels.

(a) $1-\alpha=0.99$

(b) $1-\alpha=0.95$

(c) $1-\alpha=0.90$

P5.14 Refer to Problem 5.31. For the two steel specimen types in Prob. 5.31, determine a two-sided 95% confidence interval on the true variance.

P5.15 In a random sample of 30 machine turned parts, the sample standard deviation of the diameter of the machined parts was found to be 10 mm. Compute a 90% upper confidence interval on the variance of the diameter of the turned parts.

P5.16 The following data represent the amount of corrosion coating (in ounces) need to paint 6 plates:

8.2, 8.8, 7.7, 7.9, 8.4 and 7.9

Determine the following:

(a) a point estimate of σ^2, the variance of the amount of corrosion coating needed to paint a plate

(b) confidence intervals for σ^2 and σ with the following confidence levels, assuming that the amount of corrosion coating per plate is normally distributed. *(i)* 95%, *(ii)* 98%.

P5.17 Suppose 30 samples of an allergic medicine was randomly selected and weighed. The sample standard deviation is 1.3 milligrams. Assuming the weights are normally distributed, construct a 99% confidence intervals for the population variance and standard deviation.

P5.18 The following measurements (in kg) were obtained on the weights of 6 manufactured metal parts picked at random: 12.1, 8.1, 7.2, 12.3, 14.05 and 13. Determine the confidence intervals for σ^2 and σ with 95% confidence level.

P5.19 The manufacturer of a new line of automotive transmission is estimating the fraction of non-conforming units produced. Find a 95% two-sided confidence interval for the following:

(a) 150 transmissions in the sample with 15 non-conforming

(b) 1500 transmissions in the sample with 150 non-conforming

(c) 15,000 transmissions in the sample with 1500 non-conforming

P5.20 With a random sample of 1000 people, 400 are women who turn up to see a baseball game. Find a 95% confidence interval for the percentage of women at the game.

P5.21 A sample of 600 observations selected from a population produced a sample proportion equal to 0.7.

(a) make a 90% confidence interval for *p*.

(b) construct a 95% confidence interval for *p*.

(c) make a 99% confidence interval for *p*.

(d) does the width of the confidence intervals constructed in part *(a)* through *(c)* increase as the confidence level increases?

P5.22 In a survey, 90% of drivers rated their driving as excellent or good. Suppose that this percentage was based on a random sample of 400 drivers,

(a) what is the point estimate of this corresponding population proportion?

(b) find a 95% confidence interval for the corresponding population proportion.

P5.23 A mail order company guarantees its customers that the products ordered will be mailed within 72 hours after an order is receives. The quality control department took a sample of 50 orders and found that 40 of them were mailed within 72 hours of the placement of the orders.

(a) construct a 98% confidence interval for the percentage of all orders that are mailed within 72 hours of their placement

(b) if the confidence interval found in part *(a)* is too wide, suggest a way to reduce the width of the interval.

P5.24 In a poll of 2000 adults, 60% of adults said that public education needs to be improved. Construct a 95% confidence interval for the proportion of all adults who hold this opinion.

P5.25 If a random sample of 60 non-smokers have a mean life of 75 years with a standard deviation of 7 years, and a random sample of 50 smokers live 60 years with a standard deviation of 8 years, find a 95% confidence interval for the difference of mean life time of non-smokers and smokers.

P5.26 The Chairman of the Mechanical Engineering Department is concerned about grade inflation in that department. In a random sample of 100 students in Mechanical Engineering courses this year, 80 received *A* or *B*; in a sample of 120 grades in Mechanical Engineering 5 years ago, 65 were *A* or *B*. Find a 90% confidence interval for the difference in the percentage of *A* and *B* grades now and 5 years ago.

P5.27 Machine systems for a manufacturing company are supplied by two suppliers. Machines from supplier 1 have a standard deviation of 25 units and supplier 2 have a standard deviation of 30 units. A random sample of 10 machines from supplier 1 have an average of 155 units and a similar sample of supplier 2 machines have an average of 137.4 units. Construct a 90% confidence interval on the true mean difference in the means of two machines.

P5.28 At the end of crash dieting program administered to 49 men and 36 women, the following information was obtained about the loss of weight (in kg).

	Men	Women
Mean loss of weight	$\bar{X}_1 = 16$	$\bar{X}_2 = 12$
S, sample standard deviation	$S_1 = 5$	$S_2 = 4$

Find an approximate 85% confidence interval for the difference in the mean losses.

P5.29 A consulting agency was asked by a large health insurance company to investigate if business majors were better sales persons. A sample of 40 sales persons with a business degree showed that they sold an average of 10 insurance policies per week with a standard deviations of 1.80 policies. Another sample of 45 sales persons with a degree other than business showed that they sold an average of 8.5 insurance policies per week with a standard deviation of 1.35 policies. Construct a 99% confidence interval for the difference between the two population means.

P5.30 An administrator wants to test if the mean GPAs (grade point averages) of all male and female college students who actively participate in sports are different. A random sample of 28 male and 24 female students who actively participate in sports was selected. The mean GPAs of the two groups are found to be 2.62 and 2.74 out of 4.0 respectively, with the corresponding standard deviations equal to 0.43 and 0.38. Assume that the GPAs of all male and female students have a normal distribution with the same standard deviation. Construct a 90% confidence interval for the difference between the two population means.

P5.31 Two slightly different steel specimen types of bar samples are being tested for tensile strength. Ten samples of each specimen type were selected one at random from ten different lots. The tensile tests on these bars resulted in

$$\bar{X}_1 = 25 \text{ MPa}, \; s_1^2 = 2.2 \text{ (MPa)}^2,$$

$$\bar{X}_2 = 25.5 \text{ MPa and } s_2^2 = 2.7 \text{ (MPa)}^2.$$

Assume the tensile strengths are normally distributed.

(a) find a 99% confidence interval on the difference in means

(b) if only 5 samples were selected for the second specimen type, how does this affect the confidence interval? Assume the value $\bar{X}_2$ of s_2^2 and remain unchanged.

P5.32 The outputs of two manufacturing firms over one year period with random days picked gave the following data:

$$n_1 = 20 \qquad \bar{X}_1 = 250 \qquad s_1^2 = 35$$

$$n_2 = 25 \qquad \bar{X}_2 = 200 \qquad s_2^2 = 40$$

Assuming the two outputs are normally distributed with the same variance, find a 95% two-sided confidence interval on the difference in their means.

P5.33 Random sample of weights were taken from two packages in a packaging process. The first sample consisted of $n_1 = 150$ and the second consisted of $n_2 = 250$ packages. The two sample means and standard deviations are as follows:

$$\bar{X}_1 = 10 \text{ kg} \qquad \bar{X}_2 = 9.8 \text{ kg}$$

$$s_1 = 0.25 \text{ kg} \qquad s_2 = 0.35 \text{ kg}$$

Determine the interval within which we can be 90% sure that $\mu_1 - \mu_2$ lies.

P5.34 Two diets, diet *A* and diet *B*, were fed to two groups of randomly picked cats of a certain bread. Diet *A* was fed to 10 cats. Their mean weight $\bar{X}_1$ was 10 kg, and the sample standard deviation S_1 was 3 kg. Diet *B* was fed to 14 cats. Their mean weight $\bar{X}_2$ was 12 kg and the sample standard deviation S_2 was 4 kg. Find a confidence interval for the difference in means with a 90% confidence level. Assume the weights of the cats given diet *A* and the weights of the cats given diet *B* are normally distributed with the same variance.

P5.35 Ten samples of standard cement had an average weight per cent calcium of $\bar{X}_1 = 80$, with a sample standard deviation of $S_1 = 5$, while 15 samples of lead-doped cement has an average weight per cent calcium of $\bar{X}_2 = 77$ with a sample standard deviation of $S_2 = 4$. Assume that the weight per cent calcium is normally distributed. Find a 95% confidence interval on the difference in means, $\mu_1 - \mu_2$, for two types of cement. Further, assume that both normal populations have the same standard deviation.

P5.36 A new voltage-regulating device is installed in an electromechanical system. Before the installation, a random sample yielded the following information about the percentage of breakdowns of the system: $\bar{X}_1 = 8$, $S_1^2 = 36$ and $n_1 = 8$. After installation, a random sample yields $\bar{X}_2 = 7$, $S_2^2 = 25$ and $n_2 = 9$. Find a confidence interval for the difference of the means with a 90% confidence level.

P5.37 Table P5.37 gives the systolic blood pressure of 7 adults before and after the completion of a special dietary plan based on a special dietary plan for 3 months. Construct a 95% confidence interval for μ_d. Assume that the population of paired differences, μ_d, is (approximately) normally distributed.

Table P5.37

Before	209	179	195	221	232	199	223
After	192	185	186	223	221	182	232

P5.38 A college claims that the Math tutoring service offers significantly increases the test scores of students in mathematics. The following table gives the scores per 120 of 8 students before and after they took the tutoring help.

Before	82	75	89	91	66	70	91	69
After	97	72	94	111	80	72	117	76

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the score before attending the tutoring service minus the score after attending the tutoring service. Assume that the population of paired differences is approximately normally distributed.

P5.39 A medical agency measured the corneal thickness of 8 patients who had glaucoma in one eye but not in the other. The following are the data on corneal thickness in microns.

Patient	Normal	Glaucoma
1	484	488
2	479	479
3	493	481
4	445	427
5	437	441
6	399	409
7	464	459
8	477	460

Assume that the population of paired differences is approximately normally distributed. Make a 95% confidence interval for the mean μ_d of the population difference, where a paired difference is equal to the difference in corneal thickness for normal and Glaucoma in microns.

P5.40 A company wanted to find out if attending a special course on "How to be a successful salesperson" can increase sales of its employees. The company sent 6 of its sales persons to attend this course. The following table gives the one-week sales of these salespersons before and after they attended this course.

Before	11	18	25	10	14	16
After	17	25	26	15	20	21

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the sales before and after attending the special course. Assume that the population of paired differences is approximately normally distributed.

P5.41 A company claims that its 12-week special exercise program significantly reduces weight. A random sample of 6 persons was selected, and these persons were put on this exercise program for 12 weeks. The following table gives the weight (in kg) of those 6 persons before and after the program.

Before	80	85	65	100	103	97
After	79	81	63	97	101	95

Make a 95% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the weight before joining this exercise program minus the weight at the end of the 12-week program.

P5.42 A manufacturing company sent 10 of its employees for a special training course to increase their productivity. The following table gives the one-week production of these employees before and after they attend this course.

Before	12	10	18	23	25	9	14	16	7	8
After	18	13	24	22	24	14	19	20	10	11

Make a 99% confidence interval for the mean μ_d of the population paired differences where a paired difference is equal to the production before taking this special course minus the production after taking the course. Assume the population of paired differences having a normal distribution.

P5.43 The soldering machine has a supposed life of 6 years. Two machines are under consideration for acquisition by the manufacturer. Ten samples of machine *A* were subjected to simulated operating conditions, and the sample variance of machine life was computed as $S_A^2 = 4$ years². For machine *B*, $S_B^2 = 8$ years² for $n_B = 12$. Determine a 95% two-sided confidence interval on the ratio of σ_A^2 / σ_B^2 assuming that the machine lives are normally distributed.

P5.44 The length of steel bars manufactured on two different machines is being investigated. Two random samples of sizes $n_1 = 16$ and $n_2 = 18$ are selected and the sample means and sample variances are $\bar{X}_1 = 10$, $S_1^2 = 0.4$, $\bar{X}_2 = 9.6$ and $S_2^2 = 0.45$, respectively. Assuming σ_1^2 / σ_2^2, construct a 95% two-sided confidence interval on ratio of the population variance σ_1^2 / σ_2^2. Is it reasonable to conclude that the two variances are equal?

P5.45 A study has been made to compare the aluminium contents of two brands of machine parts. Ten parts of brand *A* had an average aluminium content of 3 grams with a standard deviation of 0.4 grams, while 8 parts of brand *B* had an average aluminium content of 0.3 grams with a standard deviation of 0.5 grams. Assuming that the two sets of data are independent random samples from normal populations, with equal variance find a 98% confidence interval for σ_1^2 / σ_2^2.

P5.46 Two different grinding processes are used to finish certain machine part. Both processes can produce parts at identical mean surface roughness. The company would like to select the process having the least variability in surface roughness. A random sample of $n_1 = 11$ parts from the first process results in a sample standard deviation $S_1 = 5$ micro millimeters, and a random sample of $n_2 = 16$ parts from the second process results in a sample standard deviation of $S_2 = 4$ micro millimeters. Find a 90% confidence interval on the ratio of the two standard deviations, σ_1/σ_2.

P5.47 A study has been conducted to study the capability of a gauge by measuring the weights of 2 sheets of paper. The data are shown below:

Paper 1	$S_1 = 0.091159$	$n_1 = 15$
Paper 2	$S_2 = 0.084499$	$n_2 = 15$

Use $\alpha = 0.05$ and find a confidence interval.

P5.48 The diameter of brass rods manufactured on two different machines is being studied. Two random samples of sizes $n_1 = 15$ and $n_2 = 17$ are selected, and the sample means and sample variances are

$\bar{X}_1 = 9.75$, $S_1^2 = 0.3$, $\bar{X}_2 = 7.65$ and $S_2^2 = 0.4$ respectively. Assume that the data are drawn from a normal distribution. Construct the following:

(a) a 90% two-sided confidence interval on σ_1/σ_2

(b) a 95% two-sided confidence interval on σ_1/σ_2. Compare the width of this interval with the width of the interval in part (*a*).

(c) a 90% lower confidence bound on σ_1/σ_2

P5.49 A survey was conducted involving 500 adult whites and 300 adult blacks 18 years or older. The following question was asked: "Do you think that police officers treat criminal suspects differently in low-income neighbourhoods than in middle- or high-income neighbourhoods?". The survey resulted in 300 whites responding yes and 250 blacks responding yes.

(a) find a point estimate of $p_1 - p_2$, the difference in the proportions, where p_1 is the proportion among all white adults who would respond yes and p_2 is the proportion among all black adults who would respond yes

(b) obtain an approximate 95% confidence interval for $p_1 - p_2$.

P5.50 A company wanted to estimate the difference between the percentages of users of two electric shavers who will never switch to another brand. In a sample of 500 users of electric shaver *A* taken, 100 said that they will never switch to another brand. In another sample of 400 users of brand *B* taken, 70 said they will never switch to another brand.

(a) find the point estimate of $p_1 - p_2$, where p_1 and p_2 are the proportion of all users of brands *A* and *B* respectively.

(b) construct a 97% confidence interval for the difference between the proportions of all users of the two brands who will never switch.

P5.51 In a random sample of 90 turbine blades, 10 have a surface finish that is rougher than the specifications allow. Suppose that a modification is made in the surface roughness finishing process and that, subsequently, a second random sample of 90 turbine blades is obtained. The number of defective blades in this second sample is 8. Find an approximate 95% confidence interval on the difference in the proportion of defective turbine blades produced under the two processes.

P5.52 In a random sample of 50 people from state *A*, 40 said they favoured death penalty and 24 out of 48 from state *B* were in favour of death penalty. Find a 95% confidence interval for $p_1 - p_2$, where p_1 is the proportion of those in state *A* favouring death penalty and p_2 is the proportion of those in state *B* favouring death penalty.

P5.53 An independent health agency investigated the health of independent random samples of white and African-American elderly (aged 70 or older). Of the 5989 elderly surveyed, 629 had at least one stroke, whereas 203 of 1006 African-American elderly surveyed reported at least one stroke. Find a 95% confidence interval for the difference between the stroke incidences of white and African-American elderly.

P5.54 In a survey of 1000 drivers 25–34 years old, 27% said they buckle up, whereas 350 of 1000 drivers 45–64 years old said that they did. Find a 95% confidence interval for the difference between the proportion of seat-belt users for drivers in the age group 25–34 years and 45–64 years.

P5.55 *(a)* The standard deviation of the lifetimes of a particular brand of light bulbs is 120 hours. How large a sample must be so that we are 95% confident that the error in the estimated mean life of the lights bulbs is less than 8 hours?

(*b*) If the standard deviation of the lifetimes of the light bulbs is not known, how large the samples must be tested to be 95% confident that the error in the estimated mean life of all the light bulbs is less than 0.1σ?

P5.56 A certain manufacturing firm that manufactures a certain type of electronic part wants to estimate its mean life. Assuming that the standard deviation σ is 50 hours, find how many bulbs should be tested so as to be:

(*a*) 95% confident that the estimate $\bar{x}$ will not differ from the true mean life μ more than 10 hours

(*b*) 98% confident to accomplish accuracy in part (*a*).

Assume a normal distribution of the life of the electronic part.

P5.57 A government transportation agency wants to estimate at a 95% confidence level the mean speed for all automobiles travelling on a specific national highway. From a previous study, the agency knows that the standard deviation of speeds of automobiles travelling on this highway is 6 kilometers per hour, what sample size should the agency choose for the estimate to be within 2 kilometers per hour of the population mean?

P5.58 The school administration wants to determine a 95% confidence interval for the mean number of hours that the school students spend doing homework per week. It was known that the standard deviation for hours spent per week by all school students during homework is 10 hours. How large a sample should the school administration select so that the estimate will be within 1 hour of the population mean?

P5.59 The government would like to estimate the mean family size for all families in a particular state at a 99% confidence level. It is known that the standard deviation σ for all sizes of all families in that state is 0.7. How large a sample should the government select if it wants its estimate to be within 1% of the population mean?

P5.60 (*a*) How large a sample should be selected so that the maximum error of estimate for a 99% confidence interval for p is 0.04 when the value of the sample proportion obtained from a preliminary sample is 0.6?

(*b*) Find the most conservation sample size that will produce the maximum error for a 98% confidence interval for p equal to 0.04?

P5.61 A hardware store company guarantees all hardware deliveries within 30 minutes of the placement of orders on the plane. An agency wants to estimate the proportion of all hardware delivered within 30 minutes by the company. What is the most conservative estimate of the sample size that would limit the maximum error to within 0.03 of the population proportion for a 99% confidence interval?

P5.62 A transportation safety agency wants to estimate the proportion of all drivers who wear seat belts while driving. Assume that a preliminary study has shown that 85% of drivers wear seat belts while driving. How large should the sample size be so that the 90% confidence interval for the population proportion has a maximum error of 0.02?

P5.63 (*a*) A company wants to estimate the proportion of parts made by a particular machine that are defective. The company wants this estimate to be within 0.01 of the population proportion for a 95% confidence level. What is the most conservative estimate of the sample size that will limit the maximum error to within 0.01 of the population proportion?

(*b*) Refer to Problem P5.63(*a*). Suppose a preliminary sample of 300 parts produced by that machine showed that 6% of them are defective. How large a sample should the company select so that the 95% confidence interval for p is within 0.01 of the population proportion?

REVIEW QUESTIONS

1. Explain the meaning of the following terms used in estimation.
   - (*a*) estimator and estimate
   - (*b*) point estimate and interval estimate
2. What is the point estimator of the population mean, μ? Describe how the margin of error for a point estimate of μ is calculated?
3. What are the various alternatives for decreasing the width of a confidence interval?
4. Explain the following concepts:
   - (*a*) How the width of confidence interval decreases with an increase in the sample size?
   - (*b*) How the width of confidence interval decreases with a decrease in the confidence level?
5. Explain the difference between a confidence level and confidence interval.
6. What is the maximum error of estimate for μ for a large sample? How is it obtained?
7. What are the similarities and the differences between the standard normal distribution and the *t*-distribution?
8. Describe the parameters of a normal distribution and the *t*-distribution.
9. Explain the meaning of the degrees of freedom for a *t*-distribution.
10. What assumptions are made in using the *t*-distribution to make a confidence interval for μ?
11. What assumptions are made in using the normal distribution to make a confidence interval for the proportion population, *p*?
12. What is the point estimator of the population proportion, *p*? How is the margin of error for a point estimate of *p* calculated?
13. Describe the following terms:
   - (*a*) estimation
   - (*b*) estimator
   - (*c*) maximum error of estimate
   - (*d*) *t*-distribution

STATE TRUE OR FALSE

1. A point estimate of a parameter is the value of a statistic used to estimate the parameter. (True/False)
2. Confidence interval is an interval of numbers obtained from a point estimate of a parameter. (True/False)
3. Confidence level is the confidence we have that the parameter lies in the confidence interval (i.e., that the confidence interval contains the parameter). (True/False)
4. For a fixed sample size, increasing the confidence level improves the precision, and vice versa. (True/False)
5. A statistical procedure is said to be robust if it is sensitive to departures from the assumptions on which it is based. (True/False)
6. The margin of error equals the standard error of the mean multiplied by $z_{\alpha/2}$. (True/False)
7. The margin of error can be determined if we know only the confidence level. (True/False)
8. The confidence level can be determined if we know only the margin of error. (True/False)

9. The margin of error can be determined if we know only the confidence level, population standard deviation, and sample size. (True/False)
10. The confidence level can be determined if we know only the margin of error, population standard deviation and sample size. (True/False)
11. For samples of size *n*, the variable $t = (\bar{x} - \mu)/(s/\sqrt{n})$ has the *t*-distribution with *n* degrees of freedom. (True/False)
12. When the population standard deviation is unknown, the confidence interval procedure for a population mean cannot be based on the standardised version of $\bar{x}$. (True/False)
13. A confidence-interval estimate of a parameter consists of an interval of numbers obtained from a point estimate of the parameter and a percentage that specifies how confident we are that the parameter lies in the interval. (True/False)
14. Independent samples means that the sample selected from one of the populations has no effect or bearing on the sample selected from the other population. (True/False)
15. The pooled *t*-procedures require equal population standard deviation. (True/False)
16. The non-pooled *t*-procedures do not require equal population standard deviations. (True/False)
17. The shape of a normal distribution is determined by its standard deviation. (True/False)
18. By using a paired sample, extraneous sources of variation cannot be removed. (True/False)
19. For very small sample sizes, the normality assumption is essential for both *t*-procedures (pooled and non-pooled). (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **True** 3. **True** 4. **False** 5. **False** 6. **True** 7. **False** 8. **False** 9. **True** 10. **True**
11. **False** 12. **True** 13. **True** 14. **True** 15. **True** 16. **True** 17. **True** 18. **False** 19. **True**

CHAPTER 6

Hypothesis Testing

In Chapter 5, we examined methods for obtaining confidence intervals for means, variances and proportions. In this chapter, we present another type of statistical inference regarding these same parameters.

Quite often, there is a need to make decisions about populations on the basis of sample information. We often use inferential statistics to make decisions or judgments about the value of a parameter, such as a population mean. For example, we might need to decide whether the mean life, μ, of all machine components (such as a shaft) manufactured by a particular company differs from the advertised life of 5 years or we might want to determine whether the mean age, μ, of all automobiles has increased from the 1955 year mean of 8.5 years. In attempting to reach decisions or judgments, often it is necessary to make assumptions or guesses about the population involved. Such assumptions, which may or may not be true, are called *statistical hypothesis* and in general are statements about the probability distributions of the populations.

Decisions made about populations on the basis of sample information are called *statistical decisions*. A *hypothesis* is an assertion made about a population. That is, a hypothesis is a statement that something is true. The assertion concerns a numerical value of some parameter of the population. Procedures which helps us to decide whether to accept or reject hypothesis or to determine whether observed samples differ significantly from expected results are called *test of hypothesis*, *tests of significance* or *results of decision*.

Here, *parameters* are the statistical constants of the population (such as the mean (μ), variance (σ^2), etc.) and the statistical measures computed from the sample observations (data) are called *statistic* (such as the mean $(\overline{x})$, variance s^2, S^2, etc.).

In general, the sample size greater than or equal to 30 ($n \geq 30$) is considered as large sample in statistics whereas the sample size less than 30 ($n < 30$) is regarded as small. Nearly all the distributions can be approximated by a normal distribution.

6.1 NULL HYPOTHESIS AND ALTERNATIVE HYPOTHESIS

A *null hypothesis* is a claim (or a statement) about a population parameter that is assumed to be true until it is declared false. A null hypothesis is a hypothesis to be tested. We use the symbol H_0 to represent the null hypothesis.

An *alternative hypothesis* is a claim about a population parameter that will be true if the *null hypothesis* is false. It is a hypothesis to be considered as an alternative to the null hypothesis. We use the symbol H_1 to represent the alternative hypothesis.

The problem in a *hypothesis test* is to decide whether the null hypothesis should be rejected in favour of the alternative hypothesis. To test a hypothesis, we devise a procedure for taking a random sample, computing an appropriate *test statistic* and then rejecting or failing to reject the null hypothesis H_0. Part of this procedure is specifying the set of values for the test statistic that leads to rejection of H_0. This set of values is called the *critical region* or *rejection region* for the test.

The procedure in setting up a hypothesis test is to decide on the null hypothesis and the alternative hypothesis. For example, for specifying hypothesis test for the population mean, μ, we write

Null Hypothesis: $H_0: \mu = \mu_0$

where μ_0 is some number.

Alternative Hypothesis: The selection of the alternative hypothesis depends greatly on and should reflect the objective of the hypothesis test. Three choices are available for the alternative hypothesis. For example,

1. Whether a population mean, μ, is *different from* a specified value μ_0, we write the alternative hypothesis as

   $H_0: \mu \neq \mu_0$

   A hypothesis test whose alternative hypothesis has the above form is called a *two-tailed test*.
2. Whether a population mean, μ, is *less than* a specified value μ_0, we write the alternative hypothesis as

   $H_1: \mu < \mu_0$

   A hypothesis test whose alternative hypothesis has the above form is called a *left-tailed test*.
3. Whether a population, μ, is *greater than* a specified value μ_0, we express the alternative hypothesis as

   $H_1: \mu > \mu_0$

A hypothesis test whose alternative hypothesis has the above form is called a *right-tailed test*.

A hypothesis test is called a *one-tailed test* if it is either left-tailed or right-tailed.

If H_0 specifies the population completely, then it is known as a *simple hypothesis* otherwise, it is called *composite hypothesis*.

For instance,

$H_0: \mu = \mu_0, \sigma^2 = \sigma_0^2$ is a simple hypothesis.

On the other hand,

$H_0: \mu < \mu_0, \sigma^2 = \sigma_0^2$

$\mathrm{H}_0: \mu > \mu_0, \sigma^2 = \sigma_0^2$

$\mathrm{H}_0: \mu = \mu_0, \sigma^2 > \sigma_0^2$

$H_0: \mu = \mu_0$

$H_0: \sigma^2 = \sigma_0^2$

are all composite hypothesis.

The basic logic of hypothesis testing is as follows: Select a random sample from the given population. If the sample data are consistent with the null hypothesis, do not reject the null hypothesis; if the sample data are inconsistent with the null hypothesis (in the direction of alternative hypothesis), reject the null hypothesis and conclude that the alternative hypothesis is true.

A simple hypothesis completely determines the distribution of X, whereas a composite hypothesis does not. For instance, the simple and null hypothesis H_0: $\mu = 10$ indicates that the distribution of X is exactly the same as it was before the change, whereas the composite alternative hypothesis H_1: $\mu > 10$ is less specific. It just indicates that the new mean is greater than 10, but does not specify what exactly the new mean is. In general, the null hypothesis will be *simple* and the alternative hypothesis will be composite.

6.2 THE CRITICAL REGION

All values of the test statistic in the direction of the alternative hypothesis with a test statistic value less than or equal to the significance level α define a set called the *critical region* of the test statistic. According to the definition of α, α is the probability that the test statistic will lie in the critical region.

Example E6.1a

If H_0: $\mu = \mu_0$ is true, then the test statistic is the standard normal random variable z. Assume the level of significance α. Determine the critical regions for one-sided and two-sided alternative hypothesis.

SOLUTION:

As shown in Fig. E6.1, for a two-tailed test, the rejection region is on both the left and right. The critical values are $\pm z_{\alpha/2}$; for a left-tailed test, the critical value is $-z_{\alpha}$; and for a right-tailed test, the critical value is z_{α}. From the table in Appendix-*E*, using the standard normal table, we obtain the values of z_{α} corresponding to each of those five tail areas as shown below:

$z_{0.10}$	$z_{0.05}$	$z_{0.025}$	$z_{0.01}$	$z_{0.005}$
1.28	1.645	1.96	2.33	2.575

As an example:

For H_0: $\mu < \mu_0$, the critical region is all values $z \leq -1.28$, since $p(z \leq -1.28) = 0.10$

For H_0: $\mu > \mu_0$, the critical region is all values $z \geq 1.28$, since $p(z \geq 1.28) = 0.10$

For H_0: $\mu \neq \mu_0$, the critical region is all values $z \leq -1.645$ or $z \geq 1.645$,

since $p(z \leq -1.645) + p(z \geq 1.645) = 0.10$ or $[p(z \geq 1.645 = 0.10/2 = 0.05]$.

If the test value lies in the critical region, then the null hypothesis is rejected at the α level of significance. On the other hand, if the test value is not in the critical region, then the null hypothesis is not rejected.

6.3 TYPES OF SAMPLING ERRORS (TYPE I AND TYPE II ERRORS)

There are two important types of errors that can be made when arriving at a decision on the basis of a hypothesis test: They are

1. **Type I error:** If we reject a hypothesis when it should be accepted, we call that a *type I error* has been made. Hence, a type I error occurs when a true null hypothesis is rejected.

The value of α represents the probability of committing this type of error.

That is, the probability of this error is written as: $\alpha = P(\text{Type I Error}) = P(H_0 \mid H_0 \text{ is true})$.

The value of α represents the *significance level* of the test.

Type I error is also called a *rejection error* and Type II error is called *acceptance error.*

2. **Type II error:** A type II error occurs when a false null hypothesis is not rejected. The value of β represents the probability of making a type II error. It represents the probability that H_0 is not rejected when actually H_0 is false. The value of $(1 - \beta)$ is called the *power of the test.* It represents the probability of not making a type II error.

   That is, $\beta = P\,(\text{Type II error}) = P\,(\text{fail to reject } H_0 \mid H_0 \text{ is false})$.

   Power $= 1 - \beta = P\,(\text{reject } H_0 \mid H_0 \text{ is false})$.

The two types of errors that occur in test of hypothesis depend on each other. The values of α and β cannot be lowered simultaneously for a test of hypothesis for a fixed sample size. Lowering the value of one of them will raise the value of the other. On the other hand, one can decrease both α and β simultaneously by increasing the sample size.

The Type I and Type II errors are defined in terms of probability numbers and can be controlled to desired values. The results possible in testing a hypothesis are summarised in Table 6.1.

Table 6.1: Type I (α) error and Type II (β) error

If the decision on analysis is:	If H_0 is	
	True	False and H_1 is true
Accept H_0 Do not reject H_0	Correct decision $P = 1 - \alpha$ (probability)	Wrong decision $P = \beta$ (probability)
Reject H_0	Wrong or incorrect decision $P = \alpha$ (probability) $\Sigma P = 1.0$	Correct decision $P = 1 - \beta$ (probability) $\Sigma P = 1.0$

Summarising, we have that when a statistical test procedure based on sample data will lead to precisely one of the following four situations. Two of these situations will entail correct decisions and the other two, incorrect decisions.

1. H_0 is true and H_0 is not rejected – a correct decision.
2. H_0 is true and H_0 is rejected – an incorrect decision.
3. H_0 is false and H_0 is not rejected – an incorrect decision.
4. H_0 is false and H_0 is rejected – a correct decision.

6.4 LEVEL OF SIGNIFICANCE

In testing a given hypothesis, the maximum probability with which one would be willing or prepared to risk a Type I error is called the *level of significance* of the test. This probability often denoted by α is generally specified before any samples are drawn, so that results obtained will not influence the choice made.

The general procedure in hypothesis testing is to specify a value of the probability of type I error α, often called the *significance level* of the test, and then design the test procedure so that the probability of type II error β has a suitably small value.

6.5 TAILS OF A TEST

The rejection region for a hypothesis-testing problem can be on both sides with the non-rejection region in the middle, or it can be on the left side or on the right side of the non-rejection region. A test with two rejection regions is called a *two-tailed test*, and a test with one rejection region is called a *one-tailed test*. The one-tailed test is called a *left-tailed test* if the rejection regions is on the left tail of the distribution curve, and it is called a *right-tailed test* if the rejection region is in the right tail of the distribution curve. In such cases, the critical region is a region to one side of the distribution with area equal to the level of significance.

Table 6.2 gives critical values of z for one-tailed and two-tailed tests at various levels of significance. If under a given hypothesis, the sampling distribution of a statistic S is a normal distribution with mean μ and standard deviation σ. Then, the distribution of a standardised variable (or z-score) is given by $z = (S - \mu)/\sigma$, is the standardised normal distribution (mean 0, variance 1) as shown in Table 6.1. Table 6.3 summarises the relationships between the signs in H_0 and H_1 and the tails of a test.

Table 6.2: Critical value of *z*

Level of significance, z	0.10	0.05	0.01	0.005	0.002
Critical value of z for one-tailed test	–1.28 or 1.28	–1.645 or 1.645	–2.33 or 2.33	–2.58 or 2.58	–2.88 or 2.88
Critical value of z for two-tailed test	–1.645 or 1.645	–1.96 or 1.96	–2.58 or 2.58	–2.81 or 2.81	–3.08 or 3.08

Table 6.3: Summary of the relationships between the signs in H_0 and H_1 and the tails of a test

	Two-tailed test	Left-tailed test	Right-tailed test
Sign in the null hypothesis H_0	$=$	$=$ or $\geq$	$=$ or $\leq$
Sign in the alternative hypothesis H_1	$\neq$	$<$	$>$
Rejection region	In both tails	In the left tail	In the right tail

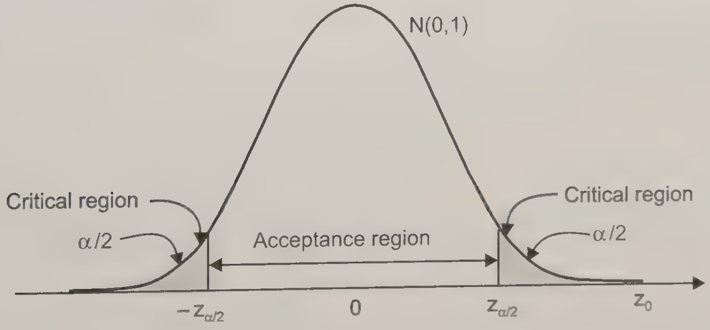


Fig. 6.1: The distribution of Z_0, when H_0: $\mu = \mu_0$ is true

For a fixed sample size, the smaller we specify the significance level, α, the larger will be the probability, β, of not rejecting a false null hypothesis. Suppose that a hypothesis test is conducted at a small significance level, then if the null hypothesis is rejected, we conclude that the alternative hypothesis is true. If the null hypothesis is not rejected, we conclude that the data do not provide sufficient evidence to support the alternative hypothesis.

To test a hypothesis, we devise a procedure for taking a random sample, computing an appropriate *test statistic*, and then rejecting or failing to reject the null hypothesis H_0. Part of this procedure is specifying the set of values for the test statistic that leads to rejection of H_0. This set of values is called the *critical region* or *rejection region* for the test.

The steps involved in performing a test of hypothesis are listed below:

1. State the null hypothesis. Under H_0 give a specific value of the parameter.
2. State the alternative hypothesis H_1. It specifies a range of possible values for the parameter that is being tested. It is important here to decide whether the test is one-tail or two-tail. Rejection of H_0 leads to acceptance of H_1.
3. Select an appropriate test statistic. The selection of the test statistic is determined by the conventional point estimator of the parameter in question.
4. Stipulate the level of significance, α, the probability of rejecting H_0 wrongly. The critical points are determined by the value of α. Formulate the decision rule based on step 2, that is, determine the values of the test statistic that will lead to the rejection of H_0 (the critical region).
5. Take a random sample and compute the value of the test statistic.
6. Make a decision based on the decision rule formulated in step 4. The choice is either reject H_0 or not reject it. Finally translate the conclusions into simple language for the benefit of the uninitiated.

6.6 HYPOTHESIS TEST ON THE POPULATION MEAN, STANDARD DEVIATION KNOWN

A basic assumptions about the population here is that it is normally distributed. In the absence of a normally distributed population, we will require that the sample size be large, that is, at least 30. The level of significance is then approximately 100 α per cent.

Suppose, the random variable X represents the population of interest. We assume that the distribution of X is either normal or that, if it is non-normal, the conditions of the central limit theorem hold.

Consider that the mean of μ of X is unknown but that the variance σ^2 is known. Suppose, we are interested in testing the hypothesis

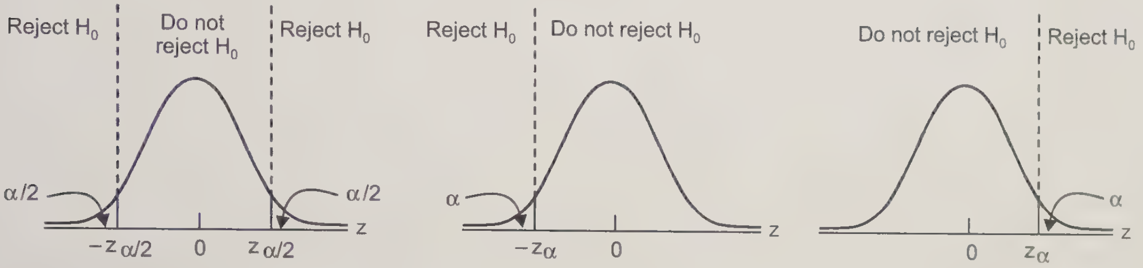
$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0 \tag{6.1}$$

where μ_0 is a specified constant.

A random sample size of n, $X_1, X_2, \ldots, X_n$, is available. Each observation in this sample has unknown mean μ and known variance σ^2. The test procedure for H_0: $\mu = \mu_0$ uses the test statistic

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \tag{6.2}$$

If the null hypothesis H_0: $\mu = \mu_0$ is true, then $E(\bar{X}) = \mu_0$, and it follows that the distribution of Z_0 is $N(0, 1)$. Consequently, if H_0: $\mu = \mu_0$ is true, the probability is $1 - \alpha$ that a value of the test statistic Z_0 falls between $-Z_{\alpha/2}$ and $Z_{\alpha/2}$, where $Z_{\alpha/2}$ is the percentage point of the standard normal distribution such that $P(Z \geq Z_{\alpha/2}) = \alpha/2$ (i.e., $Z_{\alpha/2}$) is the 100 $\alpha/2$ percentage point of the standard normal distribution). The situation is illustrated in Fig. 6.2.



(a) Two-tailed (b) Left-tailed (c) Right-tailed

Fig. 6.2

Note that the probability is α that a value of the test statistic Z_0 would fall in the region $Z_0 > Z_{\alpha/2}$ or $Z_0 < -Z_{\alpha/2}$ when H_0: $\mu = \mu_0$ is true. We should reject H_0 if either

$$Z_0 > Z_{\alpha/2} \tag{6.3}$$

or

$$Z_0 < -Z_{\alpha/2} \tag{6.4}$$

and fail to reject H_0 if

$$-Z_{\alpha/2} \leq Z_0 \leq Z_{\alpha/2} \tag{6.5}$$

Equation (6.5) defines the *acceptance region* for H_0 and Eqs. (6.4) and (6.5) defines the *critical region* or *rejection region*. The type II error probability for this test procedure is α.

A summary of the test criteria to test H_0: $\mu = \mu_0$ against the three forms of alternative hypothesis is given in Table 6.4.

Table 6.4: Procedure to test on the population mean, standard deviation known

Assumptions:
1. Simple random sample
2. Normal population or large sample
3. σ known

STEP 1: The null hypothesis is H_0: $\mu = \mu_0$, and the alternative hypothesis is

H_α: $\mu \neq \mu_0$ H_α: $\mu < \mu_0$ H_α: $\mu > \mu_0$
(Two-tailed) or (Left-tailed) or (Right-tailed)

STEP 2: Decide on the significance level, α.

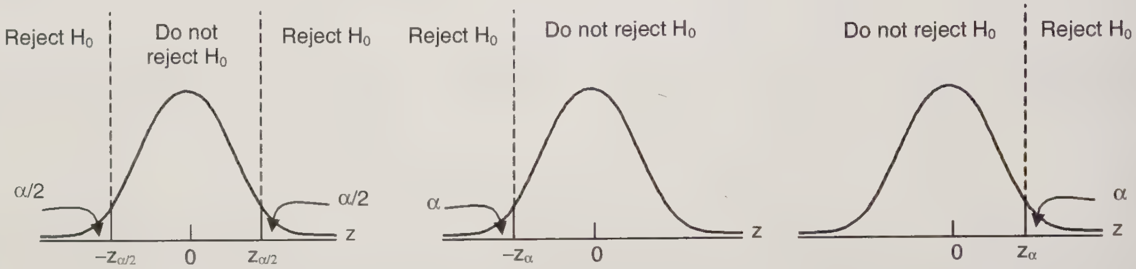
STEP 3: Compute the value of the test statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

STEP 4: The critical value(s) are

$\pm z_{\alpha/2}$ (Two-tailed) or $-z_\alpha$ (Left-tailed) or z_α (Right-tailed)

Use the table in Appendix-*E* to find the critical value(s).



(a) Two-tailed (b) Left-tailed (c) Right-tailed

STEP 5: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 6: Interpret the results of the hypothesis test.

The hypothesis test is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E6.1b

The tensile strength of steel bars produced by a manufacturer has a mean of 30 MPa and the standard deviation of 1.5 MPa. By a new technique in the manufacturing process, it is claimed that the tensile strength can be improved. To test this claim, a sample of 50 steel bars is tested and it is found that the mean tensile strength is 30.5 MPa. Can we support the claim at 0.01 level of significance?

SOLUTION:

A decision has to be made between the two hypotheses

H_0: $\mu = 30$ MPa and there is really no change in tensile strength

H_1: $\mu > 30$ Mpa and there is a change in tensile strength

A one-tailed test should be used here. At a 0.01 level of significance, the decision rule is:

1. If the z-score observed is greater than 2.33, the results are significant at the 0.01 level and H_0 is rejected. The area between 0 and z_1 in Fig. E6.1 is 0.4900, and $z_1 = 2.33$.
2. Otherwise, H_0 is accepted (or the decision is withheld). Under the hypothesis that H_0 is true, we find

$$Z_0 = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{30.5 - 30}{1.5/\sqrt{50}} = 2.357$$

which are greater than 2.33. The decision rule is: Reject H_0 if the computed value of the test statistic is greater than 2.33. Because the computed value 2.357 is greater than 2.33, we reject H_0. Hence, we can conclude that the results are "highly significant" and therefore the claim should be supported.

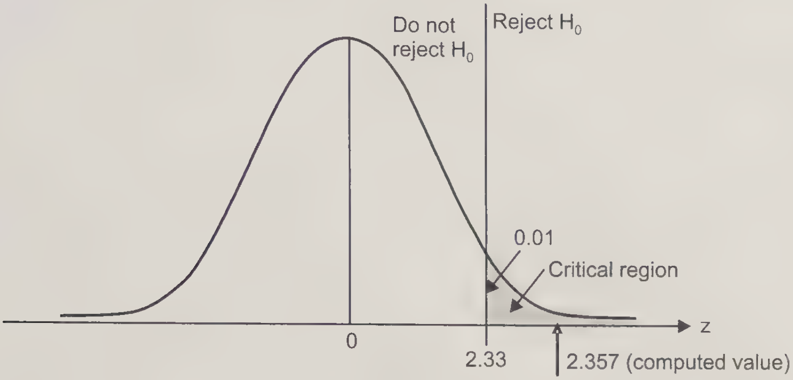


Fig. E6.1

Example E6.2

The manufacturer of a certain brand an automobile component claims that mean life of these components is 65 months. An independent agency wants to check this claim and took a random sample of 36 components and found that the mean life for this sample is 63.75 months with a standard deviation of 4 months.

(*a*) Using the 2.5% significance level, can we conclude that the mean life of these components is less than 65 months?

(*b*) Make the test a part of (*a*) using 5% significance level. Would the conclusion be different from the one in part (*a*)?

SOLUTION:

Here, a decision has to be made between two hypotheses.

H_0: $\mu = 65$ months

H_1: $\mu < 65$ months

This is a left-tailed test.

(*a*) For $\alpha = 0.025$, the critical value of z is –1.96 (From the table in Appendix-*E*).

The test statistic is

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{66.015 - 66}{0.05 / \sqrt{40}} = -1.88$$

The decision rule is: Reject H_0 if the computed value of the test statistic is less than –1.96. Because the computed value –1.88 is greater than –1.96, we do not reject H_0.

Hence, do not reject H_0 and we conclude that the mean life of these components is not less than 65 months (see Fig. E6.2(a)).

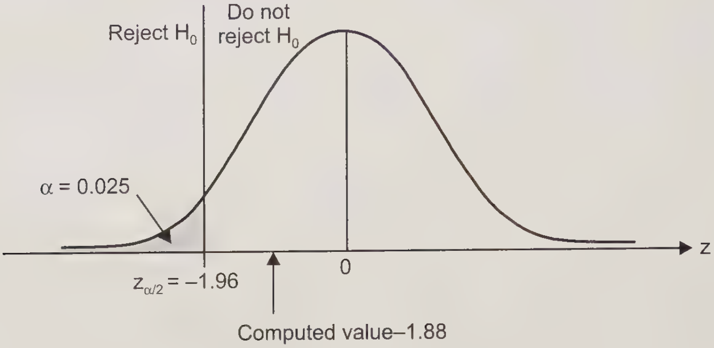


Fig. E6.2(a)

(*b*) For $\alpha = 0.05$, the critical value of $z = -1.65$ (from table in Appendix-*E*).

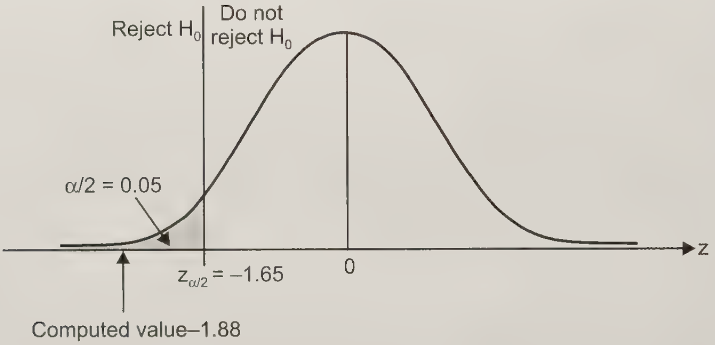


Fig. E6.2(b)

From Part (a), the value of the test statistic is –1.88. The decision rule is: Reject H_0 if the computed value of test statistic is less than –1.65. Because the computed value –1.88 is smaller than –1.65, we reject H_0. Reject H_0, and conclude that the mean life of these components is less than 65 months. The decision in parts (a) and (b) are quite different (see Fig. E6.2(b)).

Example E6.3

ABC company produces steel rods that are supposed to be 66 cm long. The mean length of the rods produced when the machine is working properly is 66 cm. The standard deviation of the length of all rods produced on this machine is 0.05 cm. A sample of 40 such steel rods was taken each week, calculated the mean length of these rods, and tested the null hypothesis against the alternative hypothesis using a 1% significance level. If the null hypothesis is rejected, the machine is stopped and adjusted. Recently a sample of 40 steel rods produced a mean length of 66.015 cm. Based on this sample, can we conclude that the machine needs an adjustment?

SOLUTION:

Here H_0: $\mu = 66$ cm

H_1: $\mu \neq 66$ cm

This is a two-tailed test.

For $\alpha = 0.01$, the critical values of z are -2.58 and 2.58 (From the table in Appendix-*E*).

The test statistic:

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \frac{66.015-66}{0.05/\sqrt{40}} = 1.90$$

The decision rule is: Reject H_0 if the computed value is less than -2.58 or greater than 2.58. Because, the computed value is less than 2.58, we do not reject H_0. Hence, do not reject H_0 and we conclude that the machine does not need to be adjusted (see Fig. E6.3).

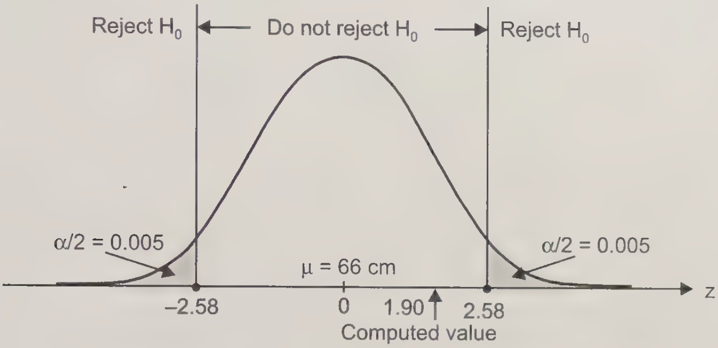


Fig. E6.3

6.7 HYPOTHESIS TEST ON THE POPULATION MEAN, STANDARD DEVIATION UNKNOWN

In tests of hypothesis regarding the population mean, the situation where the population standard deviation is not known is more realistic. The procedure presented here is particularly important when the sample size is small. The following assumptions are made:

1. The observations are independent.
2. The parent population has a normal distribution.

With the above assumptions, the statistic has students t-distribution with $n-1$ degrees of freedom, if H_0 is true.

Suppose that $X \sim N(\mu, \sigma^2)$ and wish to test H_0: $\mu = \mu_0$, and we do not known μ or σ^2. This test procedure depends on X being normally distributed, although slight departures from normally are not series, especially when the sample size is large. A random sample of size n is taken, and $\bar{X}$ and S^2 are computed.

Now, to test the two-sided hypothesis

$$H_0: \mu = \mu_0$$
$$H_1: \mu \neq \mu_0$$

The test statistic is formed,

$$t_0 = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \tag{6.5a}$$

which follows the t-distribution with $n - 1$ degrees of freedom if H_0 is true. H_0 is rejected if $|t_0| > t_{\alpha/2, v}$ where $v = n - 1$. One-sided alternatives can also be formed in a similar manner.

A summary of the test procedure to test H_0: $\mu = \mu_0$ against the three alternative hypothesis is given in Table 6.5.

Table 6.5: Procedure to test on the population mean, standard deviation known

Assumptions:
1. Simple random sample
2. Normal population or large sample
3. σ unknown

STEP 1: The null hypothesis is H_0: $\mu = \mu_0$, and the alternative hypothesis is

H_α: $\mu \neq \mu_0$ $\quad$ H_α: $\mu < \mu_0$ $\quad$ H_α: $\mu > \mu_0$

(Two-tailed) or (Left-tailed) or (Right-tailed)

STEP 2: Decide on the significance level, α.

STEP 3: Compute the value of the test statistic

$$t = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$$

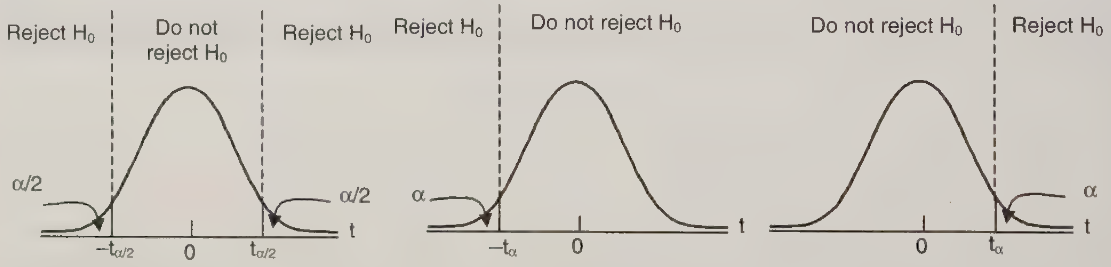
and denote that value t_0.

STEP 4: The critical value(s) are

$\pm t_{\alpha/2}$ $\quad$ $-t_\alpha$ $\quad$ t_α

(Two-tailed) or (Left-tailed) or (Right-tailed)

with $df = n - 1$. Use the table in Appendix-G to find the critical value(s).



(a) Two-tailed $\quad$ (b) Left-tailed $\quad$ (c) Right-tailed

STEP 5: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 6: Interpret the results of the hypothesis test.

The hypothesis test is exact for normal populations and is approximately correct for large samples from non-normal populations.

Example E6.4

A manufacturing company sent ten of its employees for a special training course to increase their productivity. The following table gives the one-week production of these employees before and after they attend this course.

Before	12	10	18	23	25	9	14	16	7	8
After	18	13	24	22	24	14	19	20	10	11

Using the 1% significance level, can a conclusion be made saying that "the mean weekly productions of all the employees increase as a result of attending this course"? Assume the populations of paired differences have a normal distribution.

SOLUTION:

Since the data are for paired samples, we can test the hypothesis about paired differences mean μ_0 of population using the paired differences mean $\bar{x}$ of the sample.

Let x = weekly output before the course – weekly output after the course.

Before	After	Difference	x^2
12	18	–6	36
10	13	–3	9
18	24	–6	36
23	22	1	1
25	24	1	1
9	14	–5	25
14	19	–5	25
16	20	–4	16
7	10	–3	9
8	11	–3	9

$$\sum x = -33 \text{ and } \sum x^2 = 167$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{-33}{10} = -3.3$$

$$\sigma = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}} = \sqrt{\frac{167 - \frac{(-33)^2}{10}}{10-1}} = 2.541$$

$$\frac{\sigma}{\sqrt{n}} = \frac{2.541}{\sqrt{10}} = 0.8035$$

Let $\mu_d = \mu_1 - \mu_2$

H_0: $\mu_d = 0$ ($\mu_1 - \mu_2 = 0$ or the mean output do not increase)

H_1: $\mu_d < 0$ ($\mu_1 - \mu_2 < 0$ or the mean output do increase)

Since the sample size is small (< 30), the population of paired differences is normal, and μ_d is unknown, we can use t-distribution to conduct the test. Degree of freedom = $n - 1 = 10 - 1 = 9$.

The critical value of t is obtained from the table in Appendix-G, $t_{0.01,9} = 2.821$.

The test statistic t is

$$t = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{-3.3 - 0}{0.8035} = -4.107$$

Since the test statistic $t = -4.107$ for $\bar{X}$ fall in the rejection region, we reject the null hypothesis. Therefore, we conclude that the mean weekly output for all employees increase as a result of the course.

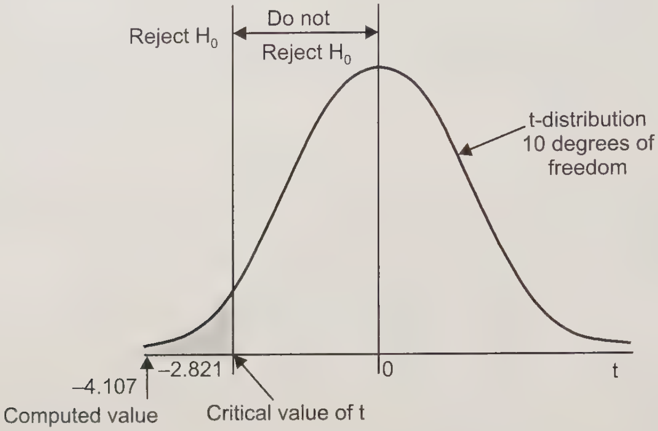


Fig. E6.4

Example E6.5

A manufacturer wants to determine whether the diameter of the steel rings gives satisfactory requirement. The particular ring will be considered satisfactory if the true diameter is greater than 20 cm. A field experiment was conducted in which 10 rings were tested at almost identical conditions and the mean diameter of the steel rings were computed for each. The results (in cm) were as follows:

23, 18, 22.1, 19.1, 19.05, 22.1, 18.2, 18.05, 23.9, and 21.9

Based on the data, is there sufficient evidence for the manufacturer to decide that the rings are satisfactory? The manufacturer is prepared to run a risk of Type I error of 0.05. It may be assumed that the diameter of the steel rings is normally distributed.

SOLUTION:

$$H_0 = \mu = 20$$

$$H_1 = \mu > 20$$

The standard deviation of the distribution is not known. We will have to estimate it from the sample. Hence, the test statistic is

$$\frac{\bar{X} - 20}{S/\sqrt{n}}$$

Here, $\alpha = 0.05$, and since $n = 10$, we have $10 - 1 = 9$ degrees of freedom. Thus, from the table for t-distribution (Appendix-G), we find $t_{9,0.05} = 1.833$. The decision rule is: Reject H_0 if the computed value of the test statistic is greater than 1.833. It can easily verified from the data that

$$\Sigma x = 205.4,\ \bar{X} = 20.54,\ \frac{\Sigma(x - \bar{x})^2}{n-1} = 5.1643$$

$$\sigma^2 = 5.1643 \text{ or } s = 2.2725$$

Hence, the computed value of the test statistic is

$$\frac{\bar{X}-20}{s/\sqrt{n}} = \frac{20.54-20}{2.2725/\sqrt{10}} = 0.7514 \text{ (see Fig. E6.5)}$$

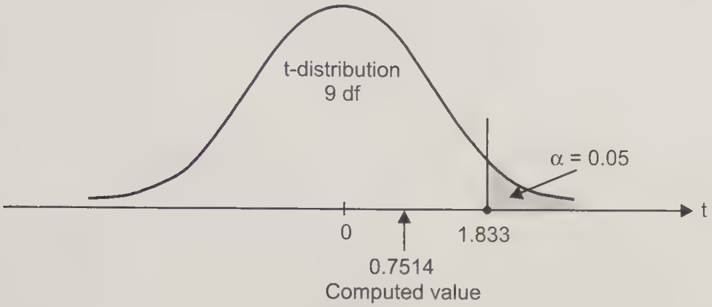


Fig. E6.5

Example E6.6

A machine can adjusted so that when under control, the mean amount of fertilizer filled in a bag is 5 kg. To check if the machine is under control, 6 bags were picked at random and their weight (in kg) were found to be as follows:

5.32, 5.19, 4.78, 5.18, 4.78 and 5.31

At the 5% level of significance, is it true that the machine is not under control? (Assume a normal distribution for the weight of a bag).

SOLUTION:

Here σ is not known and we will estimate it from the given sample.

$$H_0 = \mu = 5$$
$$H_1 = \mu \neq 5$$

The test statistic is

$$\frac{\bar{X}-5}{S/\sqrt{n}}$$

We have a two-tailed test. Given $\alpha = 0.05$ and $n = 6$, $t_{n-1,\,\alpha/2} = t_{5,0.025} = 2.571$ (from the table in Appendix-*G*). Hence, the decision rule is: Reject H_0 if the computed value is less than –2.571 or greater than 2.571.

It can easily be verified from the given sample data that

$$\bar{X} = 5.0933, \qquad s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1} = 0.0623 \text{ and } s = 0.2496$$

Thus, the computed value of the test statistic is

$$\frac{\bar{X}-20}{S/\sqrt{n}} = \frac{5.1-5}{0.2496/\sqrt{6}} = 0.9813$$

Since the computed value is between –2.571 and 2.571, we do not reject H_0.

Conclusion: There is no significant evidence at the 5% significance level to indicate that the machine is not under control, so there is no need to adjust the machine (See Fig. E6.6).

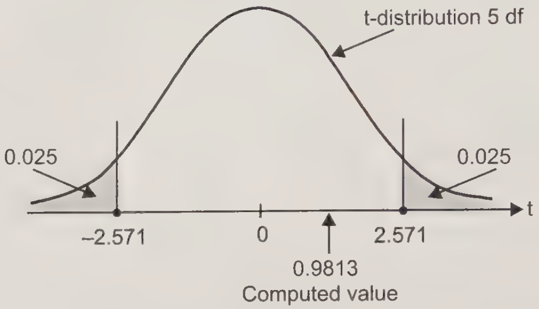


Fig. E6.6

6.8 HYPOTHESIS TEST FOR A POPULATION VARIANCE

The following assumptions are made:

1. The observations are independent.
2. The parent population has a normal distribution.

If the population is normally distributed with variance σ^2, then the distribution of $\frac{(n-1)S^2}{\sigma^2}$ is chi-square with $n - 1$ degrees of freedom, where n represents the sample size. Hence, if H_0 is true, that is, if the population variance is σ_0^2, then the distribution of $\frac{(n-1)S^2}{\sigma_0^2}$ will be chi-square with $n - 1$ degrees of freedom.

Here, we test the hypothesis that the variance of a normal distribution σ^2 equals a specified value of σ_0^2. Let $X \sim N(\mu, \sigma^2)$, where μ and σ^2 are unknown, and let $X_1, X_2, \ldots, X_n$ be a random sample of n observations from this population. To test

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 \neq \sigma_0^2 \tag{6.6}$$

We use the test statistic

$$\chi_0^2 = \frac{(n-1)S^2}{\sigma_0^2} \tag{6.7}$$

where the S^2 is the sample variance. Now if H_0: $\sigma^2 = \sigma_0^2$ is true, then the test statistic χ_0^2 follow the chi-square distribution with $n - 1$ degrees of freedom. Therefore, H_0: $\sigma^2 = \sigma_0^2$ would be rejected if

$$\chi_0^2 > \chi^2_{\alpha/2,\, n-1} \tag{6.8}$$

$$\chi_0^2 > \chi^2_{1-\alpha/2,\, n-1} \tag{6.9}$$

where $\chi^2_{\alpha/2,\, n-1}$ and $\chi^2_{1-\alpha/2,\, n-1}$ are the upper and lower $\alpha/2$ percentage point of chi-square distribution with $n - 1$ degrees of freedom.

For the one-sided hypothesis

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 > \sigma_0^2 \tag{6.10}$$

We would reject H_0 if

$$\chi_0^2 > \chi^2_{\alpha,\, n-1} \tag{6.11}$$

For the other one-sided hypothesis

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 < \sigma_0^2 \tag{6.12}$$

We would reject H_0 if

$$\chi_0^2 < \chi^2_{1-\alpha,\, n-1} \tag{6.13}$$

A summary of the test procedure to test H_0: $\sigma^2 = \sigma_0^2$ against the three alternative hypothesis is given in Table 6.6.

Table 6.6: Procedure to test for a population variance

Assumptions:

1. Simple random sample
2. Normal population

STEP 1: The null hypothesis is H_0: $\sigma^2 = \sigma_0^2$, and the alternative hypothesis is

H_a: $\sigma^2 \neq \sigma_0^2$ H_a: $\sigma^2 < \sigma_0^2$ H_a: $\sigma^2 > \sigma_0^2$

(Two-tailed) or (Left-tailed) or (Right-tailed)

STEP 2: Decide on the significance level, α.

STEP 3: Compute the value of the test statistic

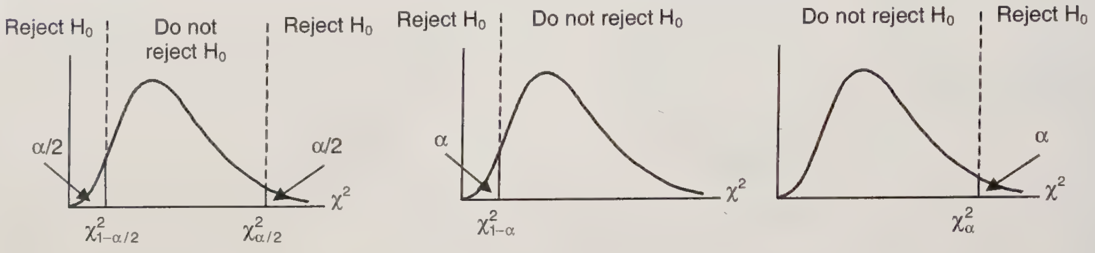
$$\chi^2 = \frac{n-1}{\sigma_0^2} S^2$$

and denote that value χ_0^2.

STEP 4: The critical value(s) are

$\chi^2_{1-\alpha/2}$ and $\chi^2_{\alpha/2}$ (Two-tailed) or $\chi^2_{1-\alpha}$ (Left-tailed) or χ^2_{α} (Right-tailed)

with $df = n - 1$. Use the table in Appendix-*F* to find the critical value(s).



Two-tailed Left-tailed Right-tailed

STEP 5: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 6: Interpret the results of the hypothesis test.

Example E6.7

The copper content in an alloy material used to make a machine part is normally distributed, and the variance is $\sigma^2 = 20$. Test the hypothesis H_0: $\sigma^2 = 20$ versus H_0: $\sigma^2 \neq 20$, if a random sample of $n = 10$ machine parts yields a sample standard deviation of $s = 4$. Use $\alpha = 0.05$.

SOLUTION:

$$\sigma^2 = 20,\ n = 10,\ s = 4$$
$$H_0\colon \sigma^2 = 20$$
$$H_0\colon \sigma^2 \neq 20$$
$$\alpha = 0.05$$

The test statistic is

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2} = \frac{(10-1)(4)^2}{20} = 7.2$$

The alternative hypothesis is two-sided so that the left is two-tailed. Now $n = 10$ and $\alpha = 0.05$.

$$\chi^2_{n-1,\alpha/2} = \chi^2_{9,0.025}$$

$$\chi^2_{n-1,1-\alpha/2} = \chi^2_{9,0.975} \quad \text{[from the table in Appendix-}F\text{]}$$

$$\chi^2_{0.05/2,n-1} = \chi^2_{0.025,9} = 19.022 \text{ [from the table in Appendix-}F\text{]}$$

$$\chi^2_{1-\alpha/2,n-1} = \chi^2_{0.975,9} = 2.70 \text{ [From the table in Appendix-}F\text{]}$$

$$\chi^2_{0.975,9} \leq \chi_0^2 \leq \chi^2_{0.025,9}$$

Do not reject H_0, the data provide insufficient evidence to claim that the variance of the copper content of the machine parts is not 20 at $\alpha = 0.05$.

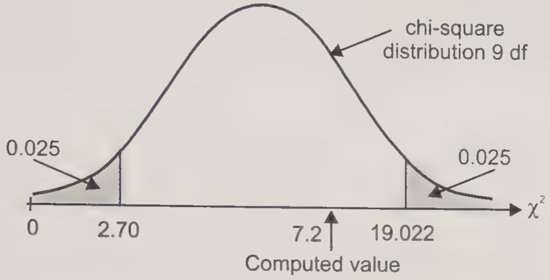


Fig. E6.7

Example E6.8

The following measurements (in kg) were obtained on the weights of 6 manufactured metal parts picked at random: 12.1, 8.1, 7.2, 12.3, 14.05 and 13. At the 5% level of significance, is the population variance of the weight of such metal parts greater than 2.3?

SOLUTION:

$$H_0: \sigma = \sqrt{2.3} = 1.5166$$

$$H_1: \sigma > 1.5166$$

The test statistic is

$$\frac{(n-1)S^2}{2.3}$$

The alternative hypothesis is right-sided. Because $n = 6$, the chi-square has 6 – 1 = 5 degrees of freedom. From the table in Appendix-*F*, $\chi^2_{5,0.05} = 11.07$. Hence, the decision rule is: Reject H_0 if the computed value of the test statistic exceeds 11.07. It can be verified that $\Sigma x = 66.75$, $\bar{X} = 11.125$ and $S^2 = 7.7916$. Therefore, the computed value of the test statistic is

$$\frac{(n-1)S^2}{2.3} = \frac{(6-1)(7.7916)}{2.3} = 16.9382 \text{ (see Fig. E6.8)}$$

Since the computed value 16.9382 is greater than 11.07, we reject H_0. We conclude that at 5% level of significance, there is evidence indicating that the population variance exceeds 2.3.

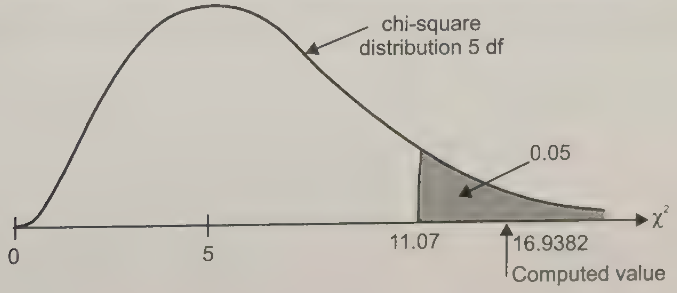


Fig. E6.8

Example E6.9

A restaurant claims that the standard deviation in the length of serving times is less than 3 minutes. A random sample of 23 serving times has a standard deviation of 2.3 minutes. At $\alpha = 0.10$, is there enough evidence to support the restaurant's claim? Assume the population is normally distributed.

SOLUTION:

The claim is "the standard deviation is less than 3 minutes".

$$H_0: \sigma \geq 3 \text{ minutes}$$

and

$$H_1: \sigma < 3 \text{ minutes}$$

The test statistic is

$$\frac{(n-1)S^2}{\sigma^2}$$

Because the test is left-tailed test, the level of significance is $\alpha = 0.10$. There are $23 - 1 = 22$ degrees of freedom and the critical value is 14.04 (from the table in Appendix-*F*). Therefore, the decision rule is: Reject H_0 if the computed value of the test statistic is less than 14.04.

$$\frac{(n-1)S^2}{\sigma^2} = \frac{(23-1)(2.3)^2}{3^2} = 12.9311 \text{ (see Fig. E6.9)}$$

Because the computed value is less than 14.04, we reject H_0.

Conclusion: There is enough evidence at the 10% level of significance to support the claim that the standard deviation for the lengths of serving times is less than 3 minutes.

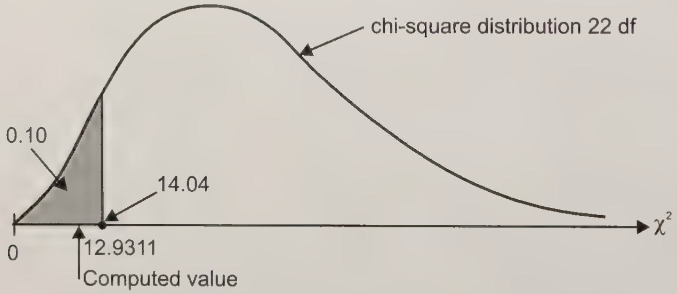


Fig. E6.9

6.9 HYPOTHESIS TEST ON A POPULATION PROPORTION

Here, we treat the case of a qualitative variable, where the data are recorded as black-white, tall-short, successes-failures, defective-nondefective, etc. The purpose will be to test hypothesis regarding the proportion p of a certain attribute in the population.

To carry out the test of hypothesis regarding the population proportion, we select a sample of observations and take the proportion of the attribute in the sample as statistic on which the test is based. We make the following assumptions:

1. The sample consists of n independent observations.
2. The sample size is large.
3. The hypothesised population proportion p_0 is not too close to 0 or 1 and is such that np_0 and $n(1 - p_0)$ are both greater than 5.

If p is the proportion in the population, then the sample proportion $\frac{X}{n}$ has a sampling distribution with mean p and standard deviation $\sqrt{p(1-p)/n}$.

If the sample size is large, both np and $n(1 - p)$ are greater than 5, then the distribution X/n is approximately normal. If the null hypothesis is true, (population proportion, p_0), then X/n will have a distribution that is approximately normal with mean p_0 and standard deviation $\sqrt{p_0(1-p)/n}$, where np_0 and $n(1 - p_0)$ are assumed greater than 5.

Consider a binomial random-variable to be tested

$$H_0: p = p_0$$
$$H_0: p^1 = p_0$$

If $0.1 \le p \le 0.9$ and n large, the procedure described is acceptable.

Pick a random sample of size n and compute $\hat{p} = X / n$, where X represents the number of observations that are nonconfirming. If H_0 is true, then $X \sim N[p_0, p_0(1-p_0)/n]$. To test $p = p_0$, calculate

$$Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \qquad (6.14)$$

and reject H_0 if $|Z_0| > Z_{\alpha/2}$. One-sided alternative could also be tested.

The above test procedure is based on the application of the central limit theorem. Therefore, the significance level is approximately 100α per cent.

Assuming of the test procedure to test H_0: $p \neq p_0$ against the three alternative hypotheses is given in Table 6.7.

Table 6.7: Procedure to test for a population proportion

Assumptions:
1. Simple random sample
2. Both np_0 and $n(1 - p_0)$ are 5 or greater

STEP 1: The null hypothesis is H_0: $p = p_0$, and the alternative hypothesis is

H_a: $p \neq p_0$	H_a: $p < p_0$	H_a: $p > p_0$
(Two-tailed) or	(Left-tailed) or	(Right-tailed)

STEP 2: Decide on the significance level, α.

STEP 3: Compute the value of the test statistic

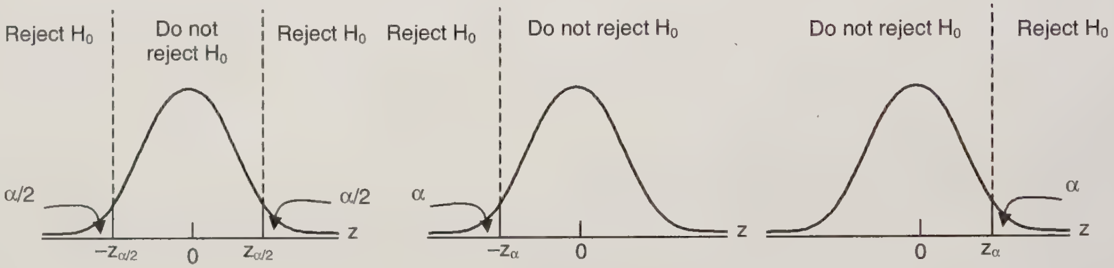
$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

and denote that value z_0.

STEP 4: The critical value(s) are

$$\pm z_{\alpha/2} \qquad -z_{\alpha} \qquad z_{\alpha}$$

(Two-tailed) or (Left-tailed) or (Right-tailed)
Use the table in Appendix-*E* to find the critical value(s).



(a) Two-tailed (b) Left-tailed (c) Right-tailed

STEP 5: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 6: Interpret the results of the hypothesis test.

Example E6.10

An internet catalogue ordering company wants to deliver on time at least 90% all the orders received from its customers. A sample of 100 orders taken for inspection showed that 83 of them were delivered on time.

(a) Using the 2% significant level, can we conclude the company's delivery policy is maintained?

(b) What will the decision be if the probability of making a type *I* error is zero?

SOLUTION:

(a)

$$H_0: p \geq 0.90$$
$$H_1: p < 0.90$$

For $\alpha = 0.02$, the critical value of Z is -2.05 (From the table in Appendix-*E*).

$$\hat{p} = \frac{83}{100} = 0.83$$

The test statistic is

$$Z = \frac{\hat{p} - p}{\sqrt{pq/n}} = \frac{0.83 - 0.90}{\sqrt{(0.90)(0.10)/100}} = -2.333$$

Because $\alpha = 0.02$, the decision rule is: Reject H_0 if the computed value is less than $-z_{0.02} = -2.05$. The computed value -2.333 is less than -2.05. Reject H_0. The conclusion is that the company's policy is not maintained.

(*b*) If $\alpha = 0$, there is no rejection region. Hence, we cannot reject H_0 and cannot conclude that the company's policy is maintained.

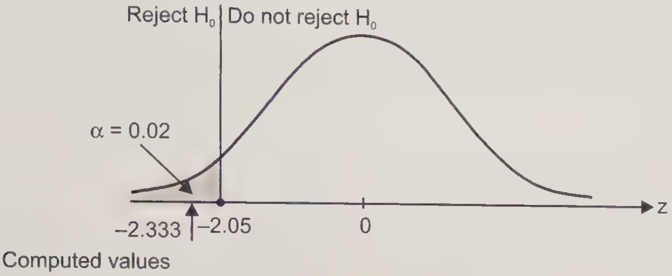


Fig. E6.10

Example E6.11

When a coin is tossed 198 times, it should heads 118 times. Is the coin biased in favour of heads? Conduct the test of hypothesis at the 5% level of significance.

SOLUTION:

Let p = probability of getting a head on a toss. Hence, we can state

$$H_0: p = 0.5 \text{ (really } p \leq 0.5)$$
$$H_1: p > 0.5$$

Since $p_0 = 0.5$, the test statistic is

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

$\alpha = 0.05$, hence the decision rule is: Reject H_0 if the computed value is greater than $z_{0.05} = 1.645$.

Since $x = 198$ and $n = 198$, the computed value of the test statistic is

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{\frac{118}{198} - 0.5}{\sqrt{(0.5)(0.5)/198}} = 2.7 \text{ (see Fig. E6.11)}$$

The computed value 2.7 is greater than 1.645. Hence, we reject H_0. We conclude that at the 5% level of significance, there is strong evidence that the coin is biased in favour of heads.

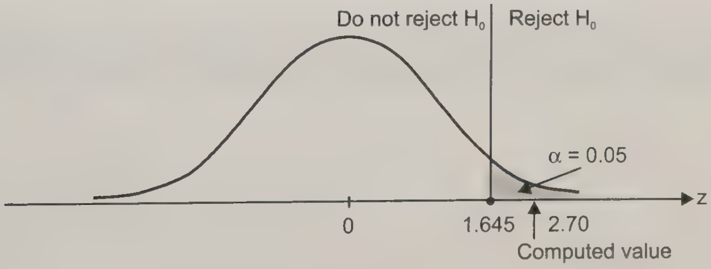


Fig. E6.11

Example E6.12

An independent survey found that 24% of the population are in favour of alcohol prohibition. You decide to test this claim and ask a random sample of 200 persons whether they are in favour of alcohol prohibition of the 200 persons, 28% are in favour. At $\alpha = 0.05$, is there enough evidence to reject the claim?

SOLUTION:

$$H_0: p = 0.24 \text{ (claim)}$$
$$H_1: p \neq 0.24$$

Because $p_0 = 0.24$, the test statistic is

$$\frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Because the test is a two-tailed test and the level of significance is $\alpha = 0.05$, the critical values are $-z_0 = -1.96$ and $z_0 = 1.96$. Therefore, the decision rule is: Reject H_0 if the computed value is less than -1.96 or greater than 1.96.

$$z = \frac{0.28 - 0.24}{\sqrt{(0.24)(0.76)/200}} = 1.324 \text{ (see Fig. E6.12)}$$

The computed value is between -1.96 and 1.96, and so we do not reject H_0.

Conclusion: At the 5% level of significance, there is not enough evidence to reject the claim that 24% of the population are in favour of alcohol prohibition.

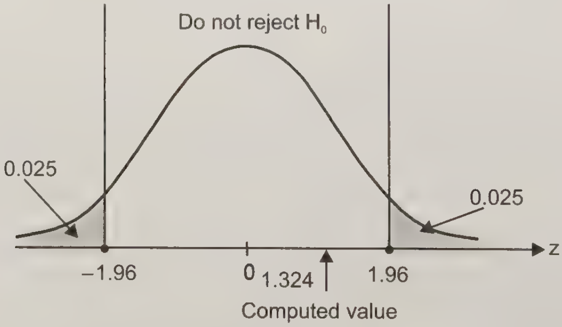


Fig. E6.12

6.10 HYPOTHESIS TEST ON EQUALITY OF TWO MEANS, VARIANCES KNOWN

There are two normal populations (or two populations from which large samples are taken) with unknown means, say μ_1 and μ_2, and known variances σ_1^2 and σ_2^2. To test the equality of the means, the following hypotheses are used:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

The test procedure consists of taking a random sample of size n_1 from the first population and compute $\bar{X}_1$. Similarly take a sample of size n_2 from the second population and compute $\bar{X}_2$. The test statistic is then given by

$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \tag{6.15}$$

with critical region $|Z_0| > Z_{\alpha/2}$.

The one-sided alternative hypotheses can also be used to test

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 > \mu_2$$

The test statistic given by Eq.(6.15) is computed, and H_0 is rejected if $Z_0 > Z_\alpha$.

To test the other one-sided alternative hypotheses

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 < \mu_2$$

The test statistic given by Eq.(6.15) is also used, but H_0 is rejected if $Z_0 < -Z_\alpha$.

The assumption of normal distribution of population is not important if sample sizes are large, that is, as a general rule, n_1 and n_2 are both at least 30. Also, here, even the population variances need not be known. The sample variance S_1^2 and S_2^2 can be used in place of respective variances σ_1^2 and σ_2^2.

Table 6.8: Procedure to test equality of two means, variances known

Assumptions:

1. Simple random sample
2. Independent samples
3. Normal populations or large samples

STEP 1: The null hypothesis is $H_0: \mu_1 = \mu_2$, and the alternative hypothesis is

$H_a: \mu_1 \neq \mu_2$	$H_a: \mu_1 < \mu_2$	$H_a: \mu_1 > \mu_2$
(Two-tailed) or	(Left-tailed)	or (Right-tailed)

STEP 2: Decide on the significance level, α.

STEP 3: Compute the value of the test statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(S_1^2 / n_1) + (S_2^2 / n_2)}}$$

Denote that value of the test statistic t_0.

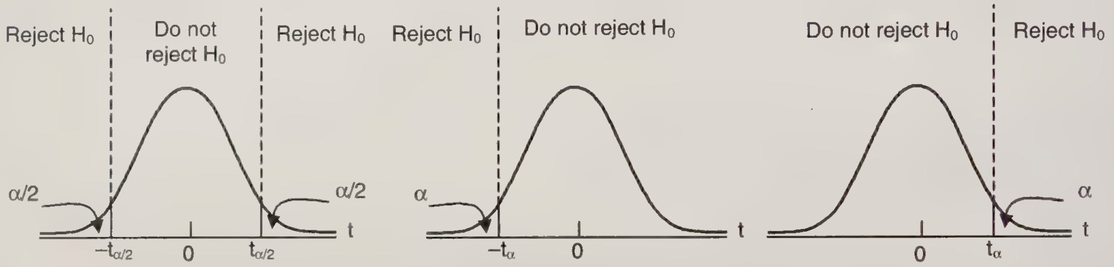
STEP 4: The critical value(s) are

$\pm t_{\alpha/2}$	$-t_\alpha$	t_α
(Two-tailed) or	(Left-tailed) or	(Right-tailed)

with $df = \Delta$, where

$$\Delta = \frac{[(S_1^2/n_1)+(S_2^2/n_2)]^2}{\dfrac{(S_1^2/n_1)^2}{n_1-1}+\dfrac{(S_2^2/n_2)^2}{n_2-1}}$$

rounded down to the nearest integer. Use the table in Appendix-*G* to find the critical value(s).



(a) Two-tailed (b) Left-tailed (c) Right-tailed

STEP 5: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 6: Interpret the results of the hypothesis test.

Example E6.13

The following is summary statistics for independent simple random samples from 2 populations. Use the nonpooled *t*-test and the nonpooled *t*-interval procedure to conduct the required hypothesis test.

Sample 1	Sample 2
$\bar{X}_1 = 7$	$\bar{X}_2 = 10.14$
$S_1 = 1.22$	$S_2 = 4.59$
$n_1 = 14$	$n_2 = 6$

Is there sufficient evidence to conclude that the mean in sample 1 exceeds that in sample 2? Assume 1% significance level.

SOLUTION:

Given $\bar{X}_1 = 115.8,\ \bar{X}_2 = 25, S_1 = 79.4, S_2 = 10.5, n_1 = 51, n_2 = 19.$

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 > \mu_2$$
$$\alpha = 0.01$$

The test statistic is

$$t = \frac{115.8 - 25}{\sqrt{(79.4^2/51)+(10.5^2/19)}} = 7.982$$

$$\Delta = \frac{\left[\left(\frac{S_1^2}{n_1}\right)+\left(\frac{S_2^2}{n_2}\right)\right]^2}{\left[\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1}\right]+\left[\frac{S_2^2}{n_2}\over n_2-1\right]^2} = \frac{\left[\left(\frac{79.4^2}{51}\right)+\left(\frac{10.5^2}{19}\right)\right]^2}{\frac{\left(\frac{79.4^2}{51}\right)^2}{(51-1)}+\frac{\left(\frac{10.5^2}{19}\right)^2}{(19-1)}} = 54.472 \approx 54$$

The critical value = 2.397 (from the table in Appendix-*G*, after interpolation).

The decision rule is: Reject H_0 if the computed value is greater than 2.397. Since the computed value 7.982 is greater than 2.397, we reject H_0 (see Fig. E6.13).

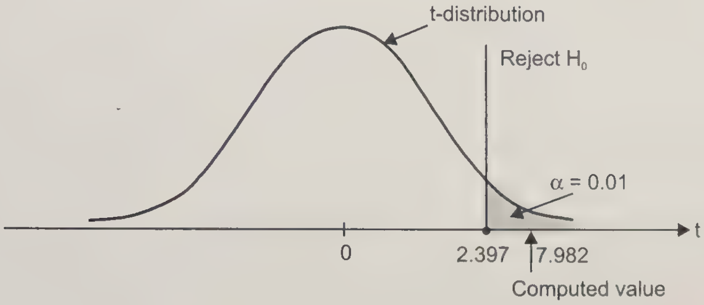


Fig. E6.13

Hence, at 1% significance level, the data do not provide sufficient evidence to conclude the mean from sample 1 exceeds that in sample 2.

Example E6.14

The following is summary statistics for independent simple random samples from 2 populations. Use the nonpooled *t*-test and the nonpooled *t*-interval procedure to conduct the required hypothesis test.

Sample 1	Sample 2
$\bar{X}_1 = 7$	$\bar{X}_2 = 10.14$
$S_1 = 1.22$	$S_2 = 4.59$
$n_1 = 14$	$n_2 = 6$

At the 5% significance level, do the data provide sufficient evidence to conclude that the mean of sample 1 smaller than sample 2?

SOLUTION:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 < \mu_2$$
$$\alpha = 0.05$$

The test is a left-tailed test.

The test statistic is

$$t = \frac{7 - 10.14}{\sqrt{\left(\frac{1.22^2}{14}\right) + \left(\frac{4.59^2}{6}\right)}} = -1.652$$

$$\Delta = \frac{\left[\left(\frac{S_1^2}{n_1}\right) + \left(\frac{S_2^2}{n_2}\right)\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left[\left(\frac{1.22^2}{14}\right) + \left(\frac{4.59^2}{6}\right)\right]^2}{\frac{\left(\frac{1.22^2}{14}\right)^2}{14 - 1} + \frac{\left(\frac{4.59^2}{6}\right)^2}{6 - 1}} = 5.306 \approx 5$$

The critical value = –2.015 (from the table in Appendix-*G*).

The decision rule is: Reject H_0 if the computed value of the test statistic is less than –2.015. Because the computed value –1.652 is greater than –2.015, we do not reject H_0 (see Fig. E6.14).

At the 5% significance level, the data do not provide sufficient evidence to conclude that the mean is smaller with sample 1 then with sample 2.

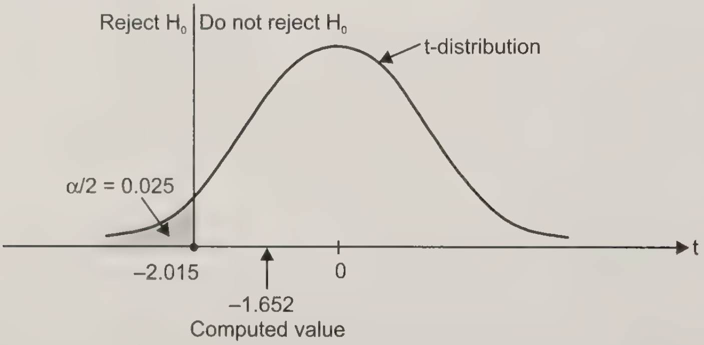


Fig. E6.14

Example E6.15

The following is summary statistics for independent simple random samples from 2 populations. Use the nonpooled *t*-test procedure to conduct the required hypothesis test.

Sample 1	Sample 2
$\bar{X}_1 = 80$	$\bar{X}_2 = 82.8$
$S_1 = 1.501$	$S_2 = 1.698$
$n_1 = 14$	$n_2 = 14$

At the 1% significance level, do the data provide sufficient evidence to conclude the means for the two samples are different?

SOLUTION:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 < \mu_2$$
$$\alpha = 0.01$$

The test statistic is

$$t = \frac{80 - 82.8}{\sqrt{\left(\frac{1.501^2}{14}\right) + \left(\frac{1.698^2}{14}\right)}} = -4.62$$

and

$$\Delta = \frac{\left[\left(\frac{S_1^2}{n_1}\right) + \left(\frac{S_2^2}{n_2}\right)\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left[\left(\frac{1.501^2}{14}\right) + \left(\frac{1.698^2}{14}\right)\right]^2}{\frac{\left(\frac{1.501^2}{14}\right)^2}{14 - 1} + \frac{\left(\frac{1.698^2}{14}\right)^2}{14 - 1}} = 25.609 \approx 25$$

The critical value = ±2.787 (from the table in Appendix-*G*).

The alternative hypothesis is two-sided so that the test is two-tailed. The decision rule is: Reject H_0 if the computed value of the test statistic is less than –2.787 or greater than 2.787. Since $-4.62 < -2.787$, we reject H_0 (see Fig. E6.15).

Conclusion: At the 1% significance level, the data provides sufficient evidence to conclude that the mean of the two samples are different.

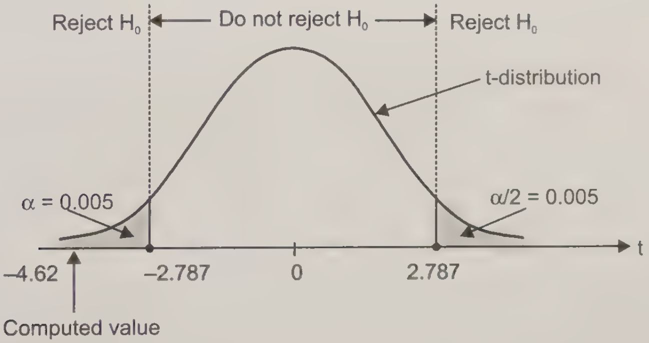


Fig. E6.15

6.11 HYPOTHESIS TEST ON THE MEANS OF TWO NORMAL DISTRIBUTIONS, VARIANCES UNKNOWN

We consider here the tests of hypotheses on the equality of the means μ_1 and μ_2 of two normally distributions where the variances σ_1^2 and σ_2^2 are unknown. A *t*-statistic is used to test the hypotheses.

6.11.1 Case 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

The following test procedures are particularly applicable for the case when small independent samples are drawn from normally distributed populations both having the same variance.

Let X_1 and X_2 be two independent normal populations with unknown means μ_1 and μ_2, and unknown but equal variances, $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We wish to test

$$\begin{aligned} H_0&: \mu_1 = \mu_2 \\ H_1&: \mu_1 \neq \mu_2 \end{aligned} \tag{6.16}$$

Suppose that $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample of n_1 observations from X_1, and $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample of n_2 observations from X_2.

Let $\bar{X}_1$, $\bar{X}_2$, S_1^2 and S_2^2 be the sample means and sample variances, respectively. S_1^2 and S_2^2 estimate the common variance σ^2 and we can combine them to yield a single estimate

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \tag{6.17}$$

To test H_0: $\mu_1 = \mu_2$ in Eq.(6.16), we compute the test statistic

$$t_0 = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{6.18}$$

If H_0: $\mu_1 = \mu_2$ is true, t_0 is distributed as $t_{n_1+n_2-2}$. Therefore, if

$$t_0 > t_{\alpha/2,\, n_1+n_2-2} \tag{6.19}$$

or if

$$t_0 < -t_{\alpha/2,\, n_1+n_2-2} \tag{6.20}$$

We reject H_0: $\mu_1 = \mu_2$.

The one-sided alternatives are treated similarly. To test

$$\begin{aligned} H_0&: \mu_1 = \mu_2 \\ H_1&: \mu_1 > \mu_2 \end{aligned} \tag{6.21}$$

compute the test statistic t_0 in Eq.(6.18) and reject H_0: $\mu_1 = \mu_2$ if

$$t_0 > t_{\alpha,\, n_1+n_2-2} \tag{6.22}$$

For the other one-sided alternative,

$$\begin{aligned} H_0&: \mu_1 = \mu_2 \\ H_1&: \mu_1 < \mu_2 \end{aligned} \tag{6.23}$$

calculate the test statistic t_0 and reject H_0: $\mu_1 = \mu_2$ if

$$t_0 < -t_{\alpha,\, n_1+n_2-2} \tag{6.24}$$

A summary of the test procedure to test H_0: $\mu_1 = \mu_2$ against the three alternative hypotheses is given in Table 6.9.

Table 6.9: Procedure to test on the means of two normal distributions, variances unknown

Assumptions:

1. Simple random sample
2. Independent samples
3. Normal populations or large samples
4. Equal population standard deviations

STEP 1: The null hypothesis is H_0: $\mu_1 = \mu_2$, and the alternative hypothesis is

H_a: $\mu_1 \neq \mu_2$	H_a: $\mu_1 < \mu_2$	H_a: $\mu_1 > \mu_2$
(Two-tailed) or	(Left-tailed)	or (Right-tailed)

STEP 2: Decide on the significance level, α.

STEP 3: Compute the value of the test statistic

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{(1/n_1)+(1/n_2)}}$$

where $$S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}.$$

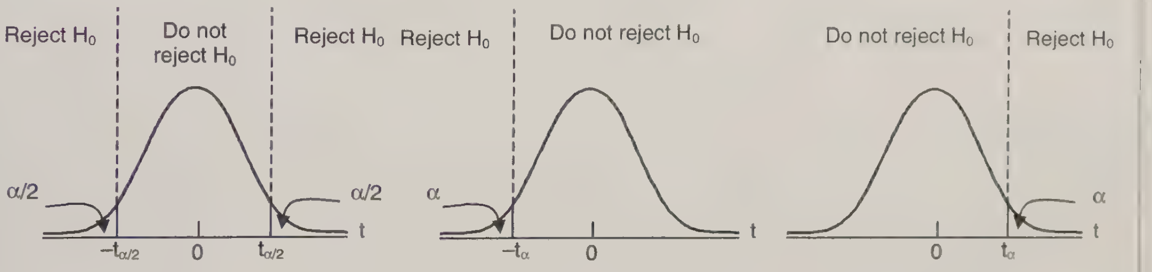
Denote that value of the test statistic t_0.

STEP 4: The critical value(s) are

$\pm t_{\alpha/2}$	$-t_\alpha$	t_α
(Two-tailed) or	(Left-tailed)	or (Right-tailed)

with $df = n_1 + n_2 - 2$.

Use the table in Appendix-*G* to find the critical value(s).



(a) Two-tailed (b) Left-tailed (c) Right-tailed

STEP 5: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 6: Interpret the results of the hypothesis test.

The hypothesis test is exact for normal populations and is approximately correct for large samples from nonnormal populations.

Example E6.16

The brands (*A* and *B*) of fertilizer were used by a company in an experiment. The following is summary statistics for the independent simple random samples from the 2 brands.

Brand	Sample size	Mean yield/plot	Standard deviation
Brand A	10	80	10
Brand B	15	88	20

At the 5% significance level, are the two brands of fertilizers significantly different? It may be assumed that the yield per plot for the two brands of fertilizers are normally distributed with the same variance.

SOLUTION:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 \neq \mu_2$$

The test statistic is

$$\frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

$$n_1 = 10, n_2 = 15,\ \bar{X}_1 = 80,\ \bar{X}_2 = 88, S_1 = 10, \text{and } S_2 = 20.$$

The degrees of freedom are $n_1 + n_2 - 2 = 10 + 15 - 2 = 23$; $\alpha = 0.05$. Therefore $t_{23,\alpha/2} = t_{23,0.025} = 2.069$ (from the table in Appendix-*G*). The decision rule is: Reject H_0 if the computed value is less than –2.069 or greater than 2.069.

$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = \frac{(10-1)(10)^2+(15-1)(20)^2}{10+15-2} = 282.61$$

or

$$S_p = 16.80$$

The test statistic is

$$\frac{80-88}{16.8\sqrt{\frac{1}{10}+\frac{1}{15}}} = -1.165 \quad \text{(see Fig. E6.16)}$$

Since the computed value –1.165 is between –2.069 and 2.069, we do not reject H_0.

We conclude that the data do not support the two brands are significantly different.

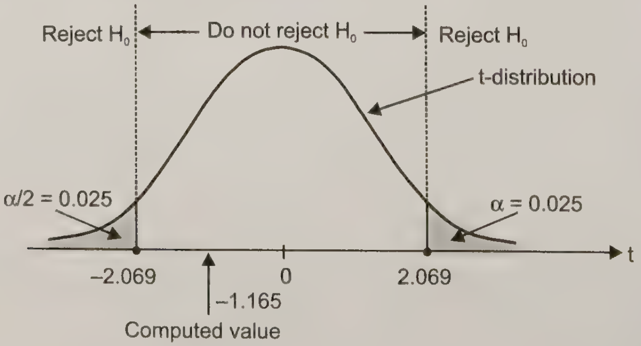


Fig. E6.16

Example E6.17

The mean life (in years) of certain machine part manufactured by two companies *A* and *B* are summarised below:

Company A	Company B
$\bar{X}_1 = 10$	$\bar{X}_2 = 18.66$
$S_1 = 4.9$	$S_2 = 4.64$
$n_1 = 10$	$n_2 = 10$

Use the pooled *t*-procedure to conduct the required hypothesis test. At the 5% significance level, do the data provide sufficient evidence to conclude that the mean life of parts made by company *A* is less than that made by machine *B*?

SOLUTION:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 < \mu_2$$
$$\alpha = 0.05$$

The test statistic is

$$\frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{10 - 18.66}{4.772\sqrt{\frac{1}{10} + \frac{1}{10}}} = -4.06$$

where $$S_p = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(10-1)(4.9)^2 + (10-1)(4.64)^2}{10 + 10 - 2} = 4.772$$

$$df = n_1 + n_2 - 2 = 10 + 10 - 2 = 18$$

The critical value is –1.734 (from the table in Appendix-*G*).

The decision rule is: Reject H_0 if the computed value of the test statistic is less than –1.734 (see Fig. E6.17). Because the computed value –4.06 is less than –1.734, we reject H_0. The data do not provide sufficient evidence to consider that, on the average, the mean life for parts made by company *A* is less than that made by Company *B*.

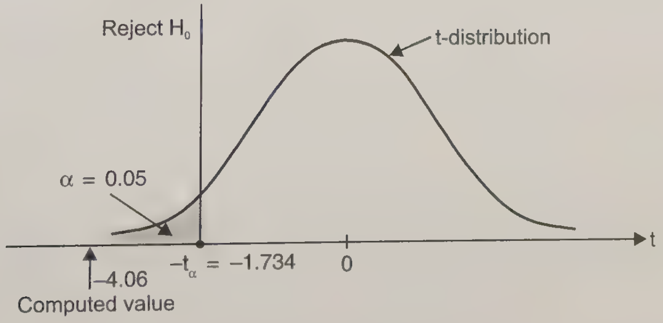


Fig. E6.17

Example E6.18

In a particular company, the mean annual salary of male workers was \$67,750 and that of female workers was \$60,000 in 1999-2000. Assume that these two means are based on random sample of 28 male and 26 female workers. The standard deviations for the two samples are \$3.7 k and \$3.2 k respectively. Using 1% significance level, can we conclude that the mean salary of all male workers in 1999–2000 was higher than that of all female workers? Further assume that the salaries of all male and female workers are both normally distributed with equal standard deviations.

SOLUTION:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 > 0$$
$$\alpha = 0.01$$

The alternative hypothesis is right-sided. The critical value of t is 2.4 (from the table in Appendix-*G* after interpolation).

$$S_p = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(28-1)(3700)^2 + (26-1)(3200)^2}{28+26-2} = 3468.6231$$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} = \frac{67{,}750 - 60{,}000}{3468.6231\sqrt{\dfrac{1}{28}+\dfrac{1}{26}}} = -8.205$$

The decision rule is: Reject H_0 if the computed value of the test statistic exceeds 2.4 (see Fig. E6.17).

Because the computed value 8.205 is greater than 2.4, we reject H_0. The mean salary of all male workers in 1999–2000 was higher than that of all female workers.

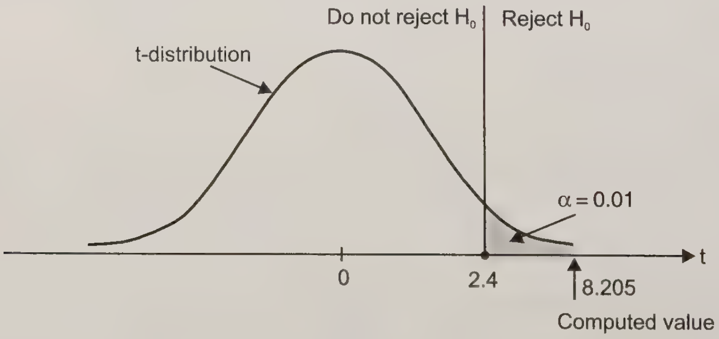


Fig. E6.18

6.11.2 Case 2: $\sigma_1^2 \neq \sigma_2^2$

In some situations, we cannot reasonably assume that the unknown variances σ_1^2 and σ_2^2 are equal. There is not an exact t-statistic available for testing H_0: $\mu_1 = \mu_2$ in this case. However, the statistic

$$t_0^* = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}}} \tag{6.25}$$

is distributed approximately as t with degrees of freedom given by

$$\nu = \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}} \tag{6.26}$$

if the null hypothesis H_0: $\mu_1 = \mu_2$ is true. Therefore, if $\sigma_1^2 \neq \sigma_2^2$, the hypothesis of Eqs. (6.16), (6.21), (6.23) are tested and t_0^* is used as the test statistic and $n_1 + n_2 - 2$ is replaced by n in determining the degrees of freedom in the test.

Example E6.19

A company claims that its high strength steel, type A provides higher tensile strength than another company's steel, type B. A researcher tested both types of steel on two groups of randomly selected ones. The results of the test are given in the following table. The mean and standard deviation of tensile strength are in MPa.

Type	Sample size	Mean of tensile strengths (MPa)	Standard deviation of tensile strength (MPa)
A	25	40	13
B	22	45	11

(a) construct a 99% confidence interval for the difference between the mean tensile strength for the two types of steel.

(b) test at a 1% significant level if the mean tensile strengths for type A is less than that for B.

SOLUTION:

(a) Degrees of freedom $= \dfrac{\left[\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right]^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$

$$\text{d.o.f.} = \frac{\left[\dfrac{13^2}{25} + \dfrac{11^2}{22}\right]^2}{\dfrac{\left(\dfrac{12^2}{25}\right)^2}{25 - 1} + \dfrac{\left(\dfrac{11^2}{22}\right)^2}{22 - 1}} = 44$$

The critical t-value (from the table in Appendix-G after interpolation), $t_{0.005,44} = 2.692$.

Hence, $$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{\frac{13^2}{25} + \frac{11^2}{22}}$$

$$(40 - 45) \pm 2.692(3.5014)$$

99% confidence interval for $\mu_1 - \mu_2$ is

$$-5 \pm 9.43 \quad \text{or} \quad -14.42 \text{ to } 4.4179$$

Type *A* is between 14.42 MPa lesser and 4.42 MPa higher than type *B*.

(*b*) H_0: $\mu_1 - \mu_2 = 0$

H_1: $\mu_1 - \mu_2 < 0$

Degrees of freedom = 44

$t_{0.01,44} = -2.414$ (from the table in Appendix-*G* after interpolation).

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(40 - 45) - 0}{3.5014} = -1.428$$

Do not reject H_0. The mean tensile strength for type *A* is not less than the mean tensile strength for type *B*.

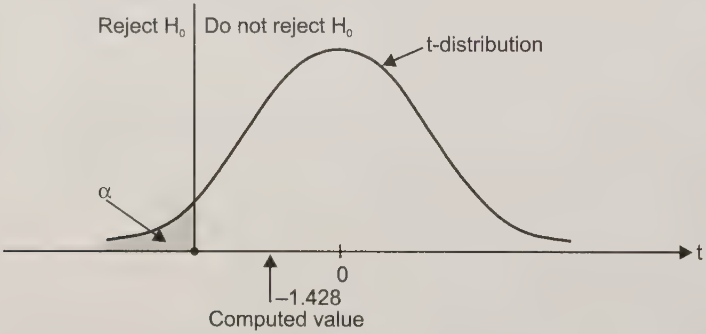


Fig. E6.19

Example E6.20

A manufacturing company is invested in buying one of two different types of machines. The company tested the two machines. The first machine was run for 8 hours and it produced an average of 129 items per hour with a standard deviation of 9 items. The second machine was run for 10 hours and it produced 120 items per hour with a standard deviation of 6 items. Assume that the production per hour for each machine is approximately normally distributed. Further assume that the standard deviations of the hourly production of the two machines are unequal.

Using the 2.5% significance level, can we conclude that the mean number of items produced per hour by the first machine is higher than that of the second machine?

SOLUTION:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 > 0$$

The appropriate statistic is

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$$

where
$$\nu = \frac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{S_1^2}{n_1}\right)^2}{n_1} + \dfrac{\left(\dfrac{S_2^2}{n_2}\right)^2}{n_2 + 1}} = 11$$

where
$$S_1 = 9, S_2 = 6, n_1 = 8 \text{ and } n_2 = 10.$$

The alternative hypothesis is one-sided (right sided). Also, $df = 11$ and $\alpha = 0.025$ so that the critical value of t is 2.201 (from the table in Appendix-*G* after interpolation).

Hence,
$$\text{test statistic} = \frac{129 - 120}{\sqrt{\dfrac{9^2}{8} + \dfrac{6^2}{10}}} = 2.429$$

Therefore, the decision rule is: Reject H_0 if the computed value of the test statistic is greater than 2.201. Because the computed value 2.429 is greater than 2.201, we reject H_0 (see Fig. E6.20).

Hence, at the 2.5% significance level, there is evidence that the mean number of items produced per hour by the first machine is higher than that by the second machine.

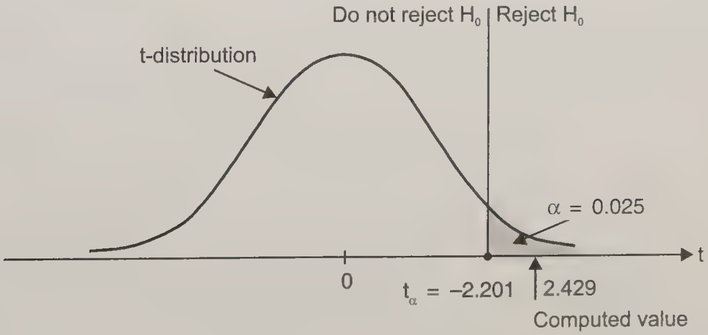


Fig. E6.20

Example E6.21

A sample of 19 cans of Brand *A* soda gave the mean number of calories per can as 15 with a standard deviation of 1.920. Another sample of 16 cans of Brand *B* soda gave the mean number of calories 18.11 per can with a standard deviation of 2.250 calories. Assume that the calories per can of soda are normally distributed for each of these brands and that the standard deviations for the two populations are not equal. Test at the 5% significance level if the mean number of calories per can of soda are different for these two brands.

SOLUTION:

Let μ_1 and μ_2 be the mean number of calories for all cans of soda of brand *A* and brand *B*, respectively. Given

$$n_1 = 8, n_2 = 10, S_1 = 9 \text{ and } S_2 = 6.$$

$$df=\frac{\left(\frac{S_1^2}{n_1}\right)+\left(\frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)}{(n_1-1)}+\frac{\left(\frac{S_2^2}{n_2}\right)}{(n_2-1)}}=\frac{\left(\frac{9^2}{8}\right)+\left(\frac{6^2}{10}\right)}{\frac{\left(\frac{9^2}{8}\right)}{(8-1)}+\frac{\left(\frac{6^2}{10}\right)}{(10-1)}}=29$$

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 > 0$$
$$df = 29$$

Area to the right tail of the *t*-curve = 0.025.

The test statistic is

$$\frac{\bar{X}_1-\bar{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}=\frac{15-18.11}{\sqrt{\frac{9^2}{8}+\frac{6^2}{10}}}=-4.353$$

The alternative hypothesis is two-sided. The critical values of *t* are ±2.045 (from the table in Appendix-*G*). Therefore, we have a two-tailed test. The decision rule is: Reject H_0 if the computed value is less than –2.045 or greater than 2.045.

Because the computed value –4.353 falls in the rejection region, we reject H_0 (see Fig. E6.21).

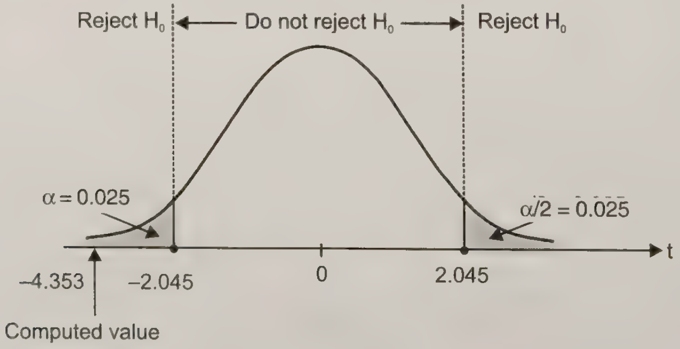


Fig. E6.21

We conclude that there is difference in the mean number of calories per can for the two brands of soda.

6.12 HYPOTHESIS TEST TO COMPARE TWO POPULATION MEANS (PAIRED t-TEST)

The method of a paired *t*-test reduces the problem of comparing the means of two populations to that of a one-sample *t*-test. Here, we denote the difference of the pair (x_i, y_i) by $d_i = x_i - y_i$.

The *n* differences $d_1, d_2, \ldots, d_n$ constitute a sample of *n* independent observations from a population of "within pair differences *D*". We assume here that the differences *D* are normally distributed (especially when *n* is small). The population mean of paired differences is then equals to $\mu_1 - \mu_2$. We are interested in testing

$$H_0: \mu_1 = \mu_2$$

This is the same as testing the null hypothesis

$$H_0: \mu_D = 0$$

Now, we have a sample of *n* independent observations $d_1, d_2, d_3, \ldots, d_n$ from a normal population whose variance is unknown, and we wish to test the hypothesis about its mean μ_D.

We can use the *t*-test with *n* – 1 degrees of freedom to test

$$H_0: \mu_D = 0$$

against the various alternative hypothesis.

The test statistic is

$$\frac{\bar{d}-0}{s_d\sqrt{n}} \text{ or } \frac{\bar{d}\sqrt{n}}{s_d}$$

where $$\bar{d} = \frac{\Sigma d_i}{n} \text{ and } s_d = \sqrt{\frac{1}{n-1}[\Sigma(d_i-\bar{d})^2]}$$

The procedure for paired differences test of a hypothesis H_0: $\mu_1 = \mu_2$ is summarised in Table 6.10.

Table 6.10: Procedure to test to compare two population means (paired t-test)

Assumptions:

1. Simple random paired sample
2. Normal differences or large sample

STEP 1: The null hypothesis is H_0: $\mu_1 = \mu_2$, and the alternative hypothesis is

H_a: $\mu_1 \neq \mu_2$ H_a: $\mu_1 < \mu_2$ H_a: $\mu_1 > \mu_2$
(Two-tailed) or (Left-tailed) or (Right-tailed)

STEP 2: Decide on the significance level, α.

STEP 3: Calculate the paired differences of the sample pairs.

STEP 4: Compute the value of the test statistic

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

and denote that value to t_0, also

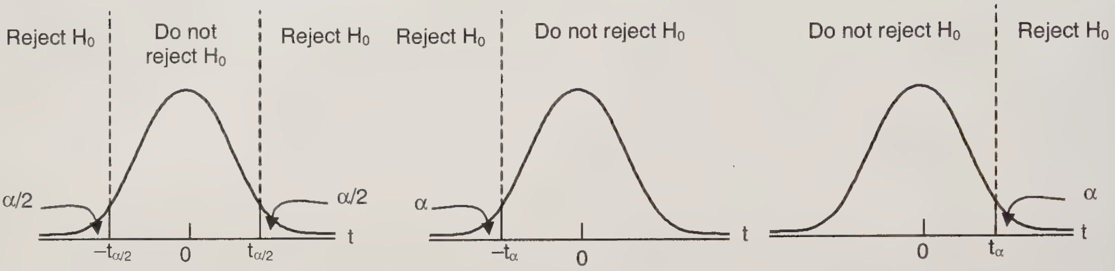
$$\bar{d} = \frac{\Sigma d_i}{n} \text{ and } s_d = \sqrt{\frac{1}{n-1}\Sigma(d_i-\bar{d})^2}$$

STEP 5: The critical value(s) are

$\pm t_{\alpha/2}$ (Two-tailed) or $-t_\alpha$ (Left-tailed) or t_α (Right-tailed)

with $df = n - 1$.

Use the table in Appendix-*G* to find the critical value(s).



(a) Two-tailed (b) Left-tailed (c) Right-tailed

STEP 6: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 7: Interpret the results of the hypothesis test.

The hypothesis test is exact for normal differences and is approximately correct for large samples and nonnormal differences.

Example E6.22

Table E6.22 gives the systolic blood pressures of seven adults and after completion of a special dietary plan based on a special dietary plan for three months. Using the 5% significance level, can we conclude that the mean of the paired differences is different from zero? Assume that the population of paired differences is approximately normally distributed.

Before	209	179	195	221	232	199	223
After	192	185	186	223	221	182	232

SOLUTION:

	Before	After	Difference, d	d^2
	209	192	17	289
	179	185	–6	36
	195	186	9	81
	221	223	–2	4
	232	221	11	121
	199	182	17	289
	223	232	–9	81
Σ	1458	1421	37	901

$$\bar{d} = \frac{\Sigma d}{n} = \frac{37}{7} = 5.2857$$

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{901 - \frac{(37)^2}{7}}{7-1}} = 10.8430$$

$$H_0: \mu_1 \neq \mu_2 \; (\mu_d = 0)$$
$$H_1: \mu_d \neq 0$$

The test statistic is

$$t = \frac{\bar{d}}{s_d \sqrt{n}} = \frac{5.2857}{10.8430\sqrt{7}} = 1.2898$$

The alternative hypothesis is two-sided. Therefore, we have a two-tailed test. $df = n - 1 = 7 - 1 = 6$ and $\alpha/2 = 0.025$.

The two critical values of t for $df = 6$ and $\alpha = 0.025$ are –2.447 and 2.447 (from the table in Appendix-*G*). The decision rule is: Reject H_0 if the computed value is less than –2.447 or greater than 2.447.

Because the computed value 1.2898 for $\bar{d}$ falls in the non-rejection region as shown in Fig. E6.22 and we fail to reject the null hypothesis.

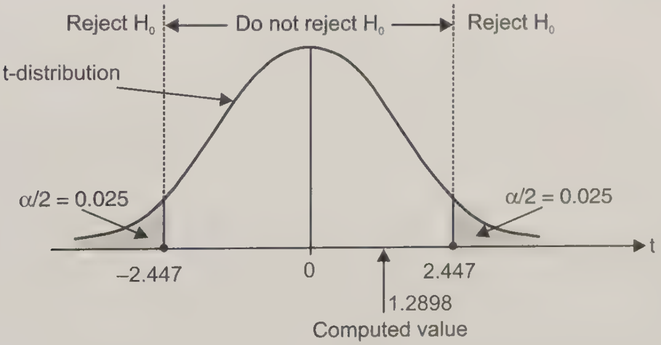


Fig. E6.22

Example E6.23

A college claims that the tutoring service it offers significantly increases the test scores of students in mathematics. The following table gives the scores out of 120 of 8 students before and after they took the tutorial help.

Before	82	75	89	91	66	70	91	69
After	97	72	94	111	80	72	117	76

Using the 5% significance level, can you conclude that taking tutorial service increases the test scores of the students? Assume that the population of paired differences is approximately normally distributed.

SOLUTION:

$$n = 8$$

	Before	After	Difference, d	d^2
	82	97	15	225
	75	72	3	9
	89	94	–5	25
	91	111	–20	400
	66	80	–14	196
	70	72	2	4
	91	117	–26	676
	69	76	–7	49
Σ	633	719	–86	1584

$$\bar{d} = \frac{\Sigma d}{n} = \frac{-86}{8} = -10.75$$

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{1584 - \frac{(86)^2}{8}}{8-1}} = 9.7064$$

$$H_0: \mu_d = 0$$

$$H_1: \mu_d < 0$$

The population statistic is

$$\frac{\bar{d}}{s_d \sqrt{n}} = \frac{-10.75}{9.7064\sqrt{8}} = -3.1325$$

The alternative hypothesis is one-sided (right sided). Also $\alpha = 0.05$. The critical value of $t = -1.895$ (from the table in Appendix-*G*).

Therefore, the decision rule is: Reject H_0 if the computed value of the test statistic is greater than –1.895.

Because the computed statistic –3.1325 is greater than –1.895, we reject H_0 (see Fig. E6.23). We conclude that attending the tutoring service increases the mean test scores.

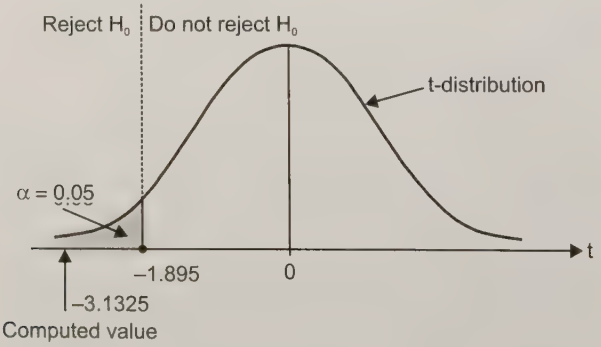


Fig. E6.23

Example E6.24

A medical agency measured the corneal thickness of 8 patients who had glaucoma in one eye but not in the other. The following are the data on corneal thickness, in microns.

Patient	Normal	Glaucoma
1	484	488
2	479	479
3	493	481
4	445	427
5	437	441
6	399	409
7	464	459
8	477	460

At the 10% significance level, do the data provide sufficient evidence to conclude the mean corneal thickness is greater in normal eyes than in eyes with glaucoma?

SOLUTION:

$n = 8$

Patient	Normal	Glaucoma	Difference, d	d^2
1	484	488	–4	16
2	479	479	0	0
3	493	481	12	144
4	445	427	18	324
5	437	441	–4	16
6	399	409	–10	100
7	464	459	5	25
8	477	460	17	289
Σ	3678	3644	34	914

$$\bar{d} = \frac{\Sigma d}{n} = \frac{34}{8} = 4.25$$

$$s_d = \sqrt{\frac{\Sigma d^2 - \frac{(\Sigma d)^2}{n}}{n-1}} = \sqrt{\frac{914 - \frac{(34)^2}{8}}{8-1}} = 10.4847$$

The test statistic is

$$t = \frac{\bar{d}}{s_d\sqrt{n}} = \frac{4.25}{10.4847\sqrt{8}} = 1.1465$$

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 > \mu_2$

$\alpha = 0.10$, $df = 7$, critical value = 1.415 (from the table in Appendix-*G*).

The alternative hypothesis is one-sided (right sided). Also, the decision rule is: Reject H_0 if the computed value of the test statistic is greater than 1.415.

Because the computed value 1.1465 is less than 1.415, do not reject H_0 (see Fig. E6.24).

Hence, the data do not provide sufficient evidence at the 10% significance level that the mean corneal thickness is greater in normal eyes than in eyes with glaucoma.

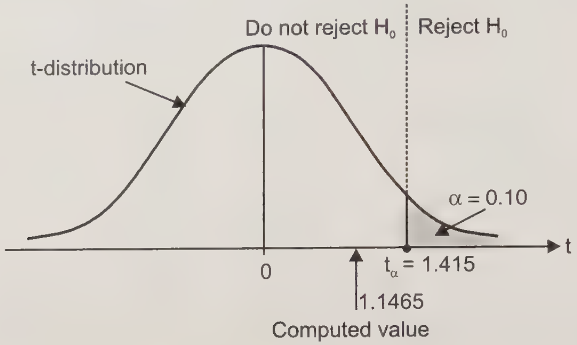


Fig. E6.24

6.13 HYPOTHESIS TEST ON THE EQUALITY OF TWO VARIANCES

The two independent populations are $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ where μ_1, σ_1^2 and μ_2, σ_2^2 are unknown. We wish to test hypotheses about the equality of the two variances, say H_0: $\sigma_1^2 = \sigma_2^2$. Assume that two random samples of size n_1 from population 1 and of size n_2 from population 2 are available, and let s_1^2 and s_2^2 be the sample variances. To test the two-sided alternative

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2 \tag{6.27}$$

The statistic

$$F_0 = \frac{S_1^2}{S_2^2} \tag{6.28}$$

is distributed as F, with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, if the null hypothesis H_0: $\sigma_1^2 = \sigma_2^2$ is true. Hence, we would reject H_0 if

$$F_0 > F_{\alpha/2, n_1-1, n_2-1} \tag{6.29}$$

or if

$$F_0 < F_{1-\alpha/2, n_1-1, n_2-1} \tag{6.30}$$

where $F_{\alpha/2, n_1-1, n_2-1}$ and $F_{1-\alpha/2, n_1-1, n_2-1}$ are the upper and lower $\alpha/2$ percentage points of the F-distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom. Appendix-G gives only the upper tail points of F, so to find $F_{1-\alpha/2, n_1-1, n_2-1}$ we must use

$$F_{1-\alpha/2, n_1-1, n_2-1} = \frac{1}{F_{\alpha/2, n_1-1, n_2-1}} \tag{6.31}$$

To test one-sided alternative hypothesis, let X_1 denotes the population that may have the largest variance, then the one-sided alternative hypothesis is

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2 \tag{6.32}$$

If

$$F_0 > F_{\alpha, n_1 - 1, n_2 - 1} \tag{6.33}$$

We would reject H_0: $\sigma_1^2 = \sigma_2^2$.

A summary of the test procedure to test H_0: $\sigma_1^2 = \sigma_2^2$ against the three alternative hypothesis is given in Table 6.11.

Table 6.11: Procedure to test on the equality of two variances

Assumptions:

1. Simple random sample
2. Independent samples
3. Normal population

STEP 1: The null hypothesis is H_0: $\sigma_1 = \sigma_2$, and the alternative hypothesis is

H_a: $\sigma_1 \neq \sigma_2$ (Two-tailed) or H_a: $\sigma_1 < \sigma_2$ (Left-tailed) or H_a: $\sigma_1 > \sigma_2$ (Right-tailed)

STEP 2: Decide on the significance level, α.

STEP 3: Compute the value of the test statistic

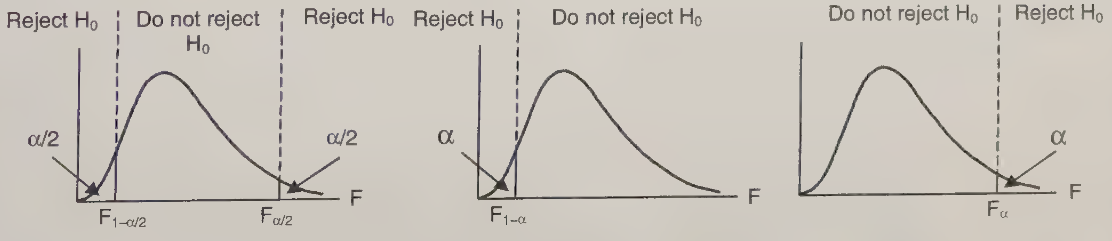
$$F = \frac{S_1^2}{S_2^2}$$

and denote that value F_0.

STEP 4: The critical value(s) are

$F_{1-\alpha/2}$ and $F_{\alpha/2}$ (Two-tailed) or $F_{1-\alpha}$ (Left-tailed) or F_{α} (Right-tailed)

with $df = (n_1 - 1, n_2 - 1)$. Use the table in Appendix-*H* to find the critical value(s).



Two-tailed   Left-tailed   Right-tailed

STEP 5: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 6: Interpret the results of the hypothesis test.

Example E6.25

A company has two branch operations, A and B, producing a particular product. Branch A has 15 employees while branch B has 25 employees on any given month, although there was no significant difference in mean production, branch A had a standard deviation of 8 while branch B had a standard deviation of 10. Can we conclude at the (a) 0.01, (b) 0.05 level of significance that the variability of branch B is greater than that of branch A?

SOLUTION:

$$n_1 = 15, n_2 = 25$$

(a) We have, on using subscripts 1 and 2 for branches A and B respectively, $S_1 = 8$, $S_2 = 10$, so that

$$\hat{S}_1^2 = \frac{n_1}{n_1 - 1} S_1^2 = \frac{15}{(15-1)}(8^2) = 68.571$$

$$\hat{S}_2^2 = \frac{n_2}{n_2 - 1} S_2^2 = \frac{25}{(25-1)}(10^2) = 104.167$$

We have to decide between the hypotheses

H_0: $\sigma_1 = \sigma_2$, and any observed variability is due to chance.

H_2: $\sigma_2 > \sigma_1$, and the variability of branch B is greater than that of A.

The decision must, therefore, be based on a one-tailed test of the F-distribution. Hence,

$$F = \frac{\hat{S}_2^2}{\hat{S}_1^2} = \frac{104.167}{68.571} = 1.519$$

$v_2 = 25 - 1 = 24$; $v_1 = 15 - 1 = 14$. At 0.01 level for 24, 14 degrees of freedom, we have from Appendix-H, $F_{0.99} = 3.43$.

Then, since $F < F_{0.99}$, we cannot reject H_0 at the 0.01 level.

(b) $F_{0.95} = 2.35$ for 24, 14 degrees of freedom (see Appendix-H).

We must note again that $F < F_{0.95}$. Thus, we cannot reject H_0 at the 0.05 level.

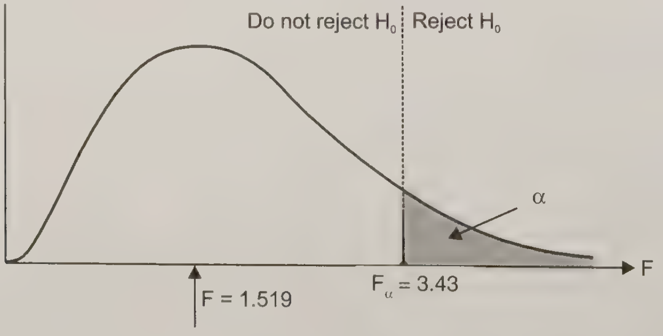


Fig. E6.25

Example E6.26

Patients who undergo chronic hemodialysis often experience severe anxiety. Videotapes of progressive relaxation exercises were shown to one group of 31 patients and neutral videotapes to another group of 25 patients. A psychiatric questionnaire was used to measure anxiety where higher scores correspond to higher anxiety. The standard deviations from the anxiety measurement were 11.0403 and 10 respectively. Do the data provide sufficient evidence to conclude that variation in anxiety-test scores differs between patients who are shown videotapes of progressive relaxation exercises and those who are shown neutral viedeotapes? Perform an *F*-test at the 10% significance level.

SOLUTION:

$$n_1 = 31 \text{ and } n_2 = 25$$

$$H_0\text{: } \sigma_1 = \sigma_2$$

$$H_0\text{: } \sigma_1 \neq \sigma_2$$

$$\alpha = 0.10$$

and test statistic

$$F = \frac{S_1^2}{S_2^2} = \frac{11.0413^2}{10^2} = 1.2189$$

The critical values are $F_{0.05}$ with df = (30, 24) or 1.94, and $F_{0.95} = \frac{1}{F_{0.05}}$ with df = (24, 30) = $\frac{1}{1.89}$ = 0.53

[see the table in Appendix-*H*].

The alternative hypothesis is two-sided. Therefore, we have a two-tailed test. The decision rule is: Reject H_0 if the computed value is less than 0.53 and greater than 1.94.

Because the computed value 1.2189 is $0.53 < 1.2189 < 1.94$, do not reject H_0 (see Fig. E6.26).

Hence, there is not sufficient evidence at the 0.10 significance level to obtain that the variation in anxiety test scores for patients seeing videotapes showing progressive relaxation exercises is different from that in patients seeing neutral videotapes

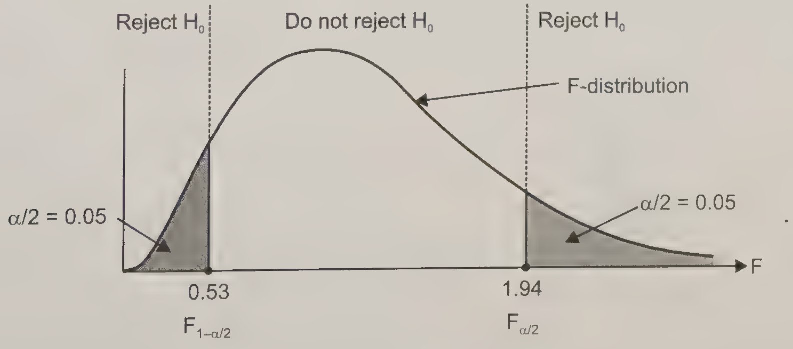


Fig. E6.26

Example E6.27

An independent golf equipment testing facility compared the differences in the performance of golf balls hit off a regular wooden tee to those hit off a brand *A* golf tee. Tests were conducted and a robot swung the club head at approximately 150 kilometer per hour. Data on ball velocity (in kilometers per hour) with each type of tee are as follows:

Brand *A*	Regular
$\bar{X}_1 = 204$	$\bar{X}_2 = 196$
$S_1 = 0.368$	$S_2 = 0.8$
$n_1 = 30$	$n_2 = 30$

At the 1% significance level, do the data provide sufficient evidence to conclude that the standard deviation of velocity is less with Brand *A* tee than with the regular tee?

SOLUTION:

$$H_0: \sigma_1 = \sigma_2$$
$$H_0: \sigma_1 < \sigma_2$$

The test statistic is

$$F = \frac{0.368^2}{0.8^2} = 0.211$$

The critical value is $F_{0.99}$ with $df = (29, 29)$ or $\frac{1}{2.42} = 0.41$ [Note that $F_{0.01} = 2.42$] from the table in Appendix-*H*. Because the alternative hypothesis is one-sided (on the left), the decision rule is: Reject H_0 if the computed value of the test statistic is less than 0.41 (see Fig. E6.27).

Because the computed value 0.211 is < 0.41, we reject the null hypothesis H_0.

There is sufficient evidence at the 0.01 significance level to claim that the standard deviation of ball velocity is less with the Brand *A* tee than with the regular tee.

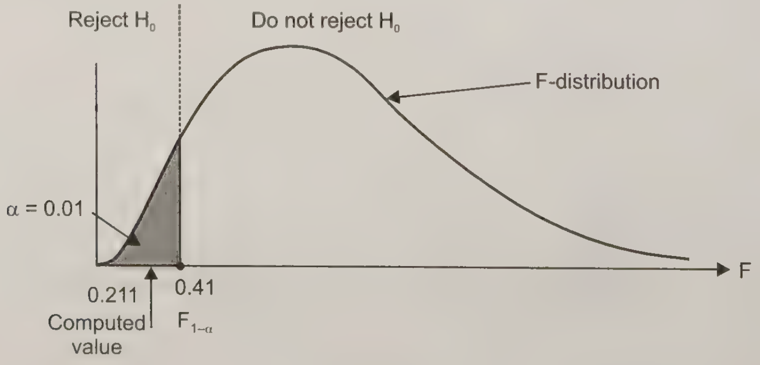


Fig. E6.27

6.14 HYPOTHESIS TEST ON TWO PROPORTIONS

Quite often, we are required to decide whether the observed differences between two sample proportions is due to sampling error or due to the fact that the proportions in the parent populations from which the samples are drawn are inherently different. Let p_1 and p_2 represent the proportions of those infected in the two populations.

The two binomial parameters are p_1 and p_2 and we with to test that they are equal. That is, H_0: $p_1 = p_2$.

If the two random samples of sizes n_1 and n_2 are taken from two populations, and X_1 and X_2 represent the number of observations that belong to the class of interest in sample 1 and 2, respectively, then assume that the normal approximation to the binomial applies to each population, so that the estimators of the population proportions $\hat{p}_1 = X_1/n_1$ and $\hat{p}_2 = X_2/n_2$ have approximate normal distribution. Now, if the null hypothesis H_0: $p_1 = p_2$ is true, then using the fact that $p_1 = p_2 = p$, the random variable

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \tag{6.34}$$

is distributed approximately $N(0,1)$. An estimate of the common parameter p is

$$\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

The test statistic for H_0: $p_1 = p_2$ is then

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} \tag{6.35}$$

If $$Z_0 > Z_{\alpha/2} \text{ or } Z_0 < -Z_{\alpha/2} \tag{6.36}$$

the null hypothesis is rejected.

A summary of the test procedure to test H_0: $p_1 = p_2$ against the three alternative hypotheses is given in Table 6.12.

Table 6.12: Procedure to test for comparison of two population proportions

Assumptions:

1. Simple random samples
2. Independent samples
3. x_1, $n_1 - x_1$, and $n_2 - x_2$ are all 5 or greater

STEP 1: The null hypothesis is H_0: $p_1 = p_2$, and the alternative hypothesis is

H_a: $p_1 \neq p_2$	H_a: $p_1 < p_2$	H_a: $p_1 > p_2$
(Two-tailed) or	(Left-tailed) or	(Right-tailed)

STEP 2: Decide on the significance level, α.

STEP 3: Compute the value of the test statistic

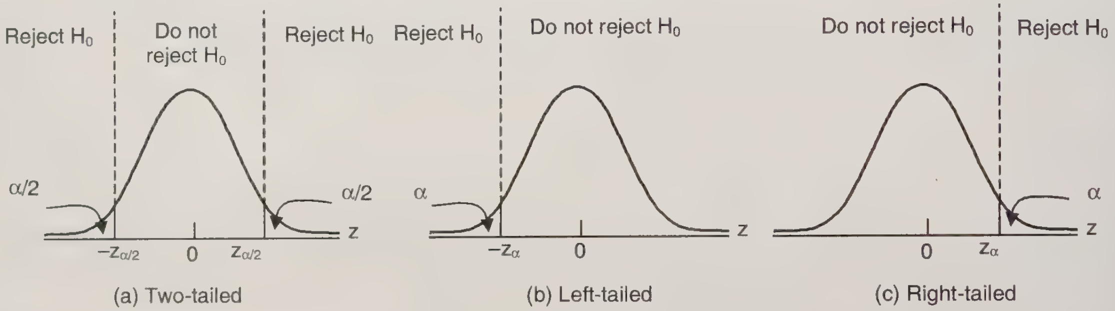
$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{(1/n_1) + (1/n_2)}}$$

where $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$. Denote the value of the test statistic z_0.

STEP 4: The critical value(s) are

$\pm z_{\alpha/2}$ (Two-tailed) or $-z_\alpha$ (Left-tailed) or z_α (Right-tailed)

Use the table in Appendix-*E* to find the critical value(s).



(a) Two-tailed (b) Left-tailed (c) Right-tailed

STEP 5: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

STEP 6: Interpret the results of the hypothesis test.

Example E6.28

In an independent study it was found that a sample of 500 male registered voters showed that 60% of them voted in the last general elections. Another sample of 400 female registered voters showed that 50% of them voted in the same election. Test at the 1% significance level if all male voters who voted in the last general election is different from that of all female voters.

SOLUTION:

$$H_0: p_1 - p_2 = 0$$
$$H_1: p_1 - p_2 \neq 0$$

$$\hat{p}_1 = 0.60,\ \hat{p}_2 = 0.50,\ n_1 = 500 \text{ and } n_2 = 400$$

For $\alpha = 0.01$, the critical values of z are -2.58 and 2.58 (from the table in Appendix-*E*).

$$\hat{p} = \frac{500(0.60) + 400(0.50)}{500 + 400} = 0.5556,\ \hat{p} = 1 - \hat{p} = 1 - 0.5556 = 0.4444$$

$$\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{(0.5556)(0.4444)\left(\frac{1}{500} + \frac{1}{400}\right)} = 0.03333$$

Hence
$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{0.60 - 0.50}{0.03333} = 3.0$$

Hence, we reject H_0 and conclude that the proportion of all male voters who voted in the last general elections is different from that of all female voters who voted in the same election (see Fig. E6.28).

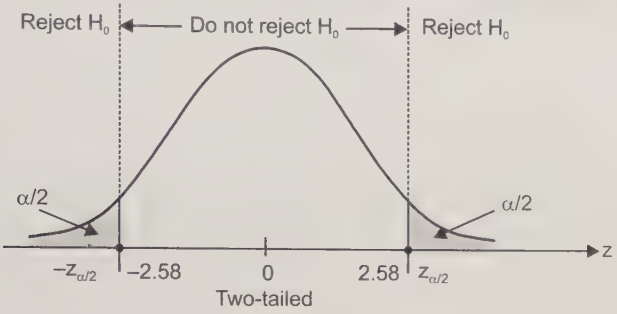


Fig. E6.28

Example E6.29

A survey found that 44% male students and 31% female students in a particular college had more than one hour of physical exercise time per day. Suppose that these percentages are based on a random sample of 395 male students and 410 female students. At the 1% significance level, can we conclude that the percentage of all male students and of all female students who have more than one hour of physical exercise time per day are different?

SOLUTION:

Here
$$H_0: p_1 - p_2 = 0$$
$$H_1: p_1 - p_2 \neq 0$$

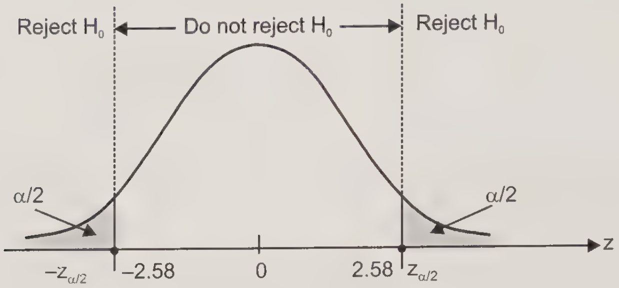
For $\alpha = 0.01$, the critical values of z are –2.58 and 2.58 (from the table in Appendix-*E*).

$$\hat{p}_1 = 0.44,\ \hat{p}_2 = 0.31,\ n_1 = 395, \text{ and } n_2 = 410$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\hat{p}(1-\hat{p})}} = \frac{(0.44 - 0.31) - 0}{\sqrt{(0.374)(0.626)\left(\frac{1}{395} + \frac{1}{410}\right)}} = 3.81$$

where
$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{0.44(395) + 0.31(410)}{395 + 410} = 0.374$$

Hence, $z = 3.81$ and we reject H_0 and conclude that the percentage of all male students and all female students who have more than one hour of physical exercise time per day are different as shown in Fig. E6.29.



Two-tailed

Fig. E6.29

Example E6.30

An independent medical research team conducted a study to test the effect of a cholesterol reducing medication. At the end of study, the team found that of the 9400 subjects who took the medication, 602 died of heart disease. Of the 8600 persons who took brand *A*, 714 died of heart disease. At $\alpha = 0.01$, can we conclude that the death rate is lower for those who took the medication than for those who took the brand *A*?

SOLUTION:

$$H_0: p_1 \geq p_2$$
$$H_1: p_1 < p_2 \text{ (claim)}$$

Because the test is left-tailed and the level of significance is $\alpha = 0.01$, the critical value is –2.33 (from the table in Appendix-*E*). The decision rule is: Reject H_0 if the computed value of the test statistic is less than –2.33.

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{602}{9400} = 0.064$$

$$\hat{p}_2 = \frac{x_2}{n_2} = \frac{714}{8600} = 0.083$$

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{602 + 714}{9400 + 8600} = 0.073$$

$$\hat{q} = 1 - \hat{p} = 1 - 0.073 = 0.927$$

The test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.064 - 0.083}{\sqrt{(0.073)(0.927)\left(\frac{1}{9400} + \frac{1}{8600}\right)}} = -4.8989$$

Because the computed value –4.8989 is less than –2.33, we reject H_0 (see Fig. E6.30). At 1% level, there is enough evidence to conclude that the death rate is lower for those who took the medication than for those who took the brand *A*.

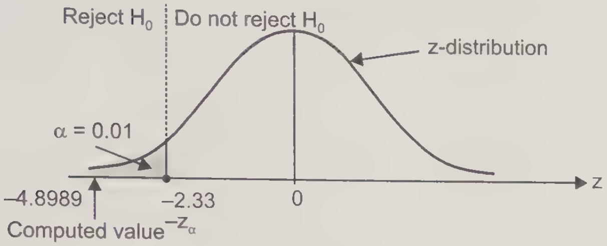


Fig. E6.30

Example E6.31

A survey of 872 workers showed that 384 of them said that it was seriously unethical to monitor employee e-mail. When 242 senior-level managers were surveyed, 80 of them said that it was seriously unethical to monitor employee e-mail. Use a 0.05 significance level to test the claim that for those saying monitoring e-mail is seriously unethical, the proportion of employee is greater than the proportion of managers.

SOLUTION:

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{384}{872} = 0.4404$$

$$\hat{p}_2 = \frac{x_2}{n_2} = \frac{80}{242} = 0.3306$$

$$\hat{p} = \frac{384+80}{872+242} = 0.4165$$

$$H_0: p_1 - p_2 = 0$$
$$H_1: p_1 - p_2 > 0$$
$$\alpha = 0.05$$

The critical values of $z_\alpha = 1.645$ (from the table in Appendix-*E*).

The alternative hypothesis is one-sided (right-side). Therefore, the decision rule is Reject H_0 if the computed value of the test statistic is greater than 1.645.

The test statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{0.4404 - 0.3306}{\sqrt{0.4165(0.5835)\left(\frac{1}{872}+\frac{1}{242}\right)}} = 3.0654$$

Because the computed value 3.0654 is greater than 1.645, we reject H_0 (see Fig. E6.31).

Hence, there is sufficient evidence to conclude that $p_1 - p_2 > 0$.

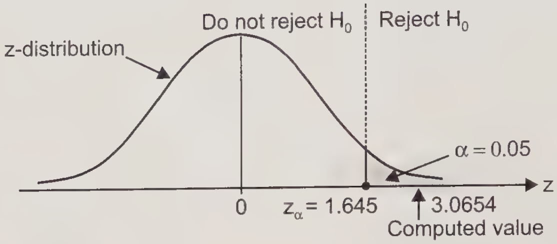


Fig. E6.31

Procedures for testing hypotheses on means, variances, and proportions for a single population and two populations are summarised in Tables 6.13 and 6.14 respectively.

Table 6.13: Summary of tests of hypotheses (single population)

No.	Nature of the population	Parameter	Null hypothesis	Alternative hypothesis	Test statistic	Decision rule: Reject the computed value
1.	Quantitative data; variance σ^2 is known; population normally distributed	μ	$\mu = \mu_0$	1. $\mu > \mu_0$ 2. $\mu < \mu_0$ 3. $\mu \neq \mu_0$	$\dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$	1. greater than z_α 2. less than $-z_\alpha$ 3. less than $-z_{\alpha/2}$ or greater than $z_{\alpha/2}$
2.	Quantitative data; variance σ^2 is not known; population normally distributed; sample size small $(n < 30)$	μ	$\mu = \mu_0$	1. $\mu > \mu_0$ 2. $\mu < \mu_0$ 3. $\mu \neq \mu_0$	$\dfrac{\bar{X} - \mu_0}{S / \sqrt{n}}$	1. greater than $t_{n-1,\alpha}$ 2. less than $- t_{n-1,\alpha}$ 3. less than $- t_{n-1,\alpha/2}$ or greater than $t_{n-1,\alpha/2}$
3.	Quantitative data; variance σ^2 is not known; population not necessarily normal; sample large $(n \geq 30)$	μ	$\mu = \mu_0$	1. $\mu > \mu_0$ 2. $\mu < \mu_0$ 3. $\mu \neq \mu_0$	$\dfrac{\bar{X} - \mu_0}{S / \sqrt{n}}$	1. greater than z_α 2. less than $-z_\alpha$ 3. less than $-z_{\alpha/2}$ or greater than $z_{\alpha/2}$ The level of significance approximately 100α per cent
4.	Quantitative data; population normally distributed	σ	$\sigma = \sigma_0$	1. $\sigma > \sigma_0$ 2. $\sigma < \sigma_0$ 3. $\sigma \neq \sigma_0$	$\dfrac{(n-1)S^2}{\sigma_0^2}$	1. greater than $\chi^2_{n-1,\alpha}$ 2. less than $\chi^2_{n-1,1-\alpha}$ 3. less than $\chi^2_{n-1,1-\alpha/2}$ or greater than $\chi^2_{n-1,\alpha/2}$
5.	Quantitative data; binomial case	p	$p = p_0$	1. $p > p_0$ 2. $p < p_0$ 3. $p \neq p_0$	$\dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$ Here n is assumed large.	1. greater than z_α 2. less than $-z_\alpha$ 3. less than $-z_{\alpha/2}$ or greater than $z_{\alpha/2}$ The level of significance approximately 100α per cent

Table 6.14: Summary of tests of hypotheses (two populations)

No.	Nature of the population	Parameter	Null hypothesis	Alternative hypothesis	Test statistic	Decision rule: Reject H_0 if the computed value is
1.	Quantitative data; variances σ_1^2 and σ_2^2 are known; both populations normally distributed	$\mu_1 - \mu_2$	$\mu_1 = \mu_2$	1. $\mu_1 > \mu_2$ 2. $\mu_1 < \mu_2$ 3. $\mu_1 \neq \mu_2$	$\dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	1. greater than z_α 2. less than $-z_\alpha$ 3. less than $-z_{\alpha/2}$ or greater than $z_{\alpha/2}$
2.	Quantitative data; σ_1^2, σ_2^2 unknown; populations not necessarily normal; sample sizes are large	$\mu_1 - \mu_2$	$\mu_1 = \mu_2$	1. $\mu_1 > \mu_2$ 2. $\mu_1 < \mu_2$ 3. $\mu_1 \neq \mu_2$	$\dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$	1. greater than z_α 2. less than $-z_\alpha$ 3. less than $-z_{\alpha/2}$ or greater than $z_{\alpha/2}$
3.	Quantitative data; σ_1^2 and σ_2^2 are not known, assumed to be equal; Both populations are normally distributed.	$\mu_1 - \mu_2$	$\mu_1 = \mu_2$	1. $\mu_1 > \mu_2$ 2. $\mu_1 < \mu_2$ 3. $\mu_1 \neq \mu_2$	$\dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ $df = \dfrac{\left[\left(S_1^2/n_1\right) + \left(S_2^2/n_2\right)\right]^2}{\dfrac{(S_1^2/n_1)^2}{n_1 - 1} + \dfrac{(S_2^2/n_2)^2}{n_2 - 1}}$	1. greater than $t_{m+n-2,\alpha}$ 2. less than $-t_{m+n-2,\alpha}$ 3. less than $-t_{m+n-2,\alpha/2}$ or greater than $t_{m+n-2,\alpha/2}$
4.	Quantitative data; σ_1^2 and σ_2^2 are not known but are assumed to be equal; both populations are normally distributed.	$\mu_1 - \mu_2$	$\mu_1 = \mu_2$	1. $\mu_1 > \mu_2$ 2. $\mu_1 < \mu_2$ 3. $\mu_1 \neq \mu_2$	$\dfrac{\bar{X}_1 - \bar{X}_2}{S_p\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ where $S_p = \sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$	1. greater than $t_{m+n-2,\alpha}$ 2. less than $-t_{m+n-2,\alpha}$ 3. less than $-t_{m+n-2,\alpha/2}$ or greater than $t_{m+n-2,\alpha/2}$

5.	Paired observations	$\mu_1 - \mu_2$ $(=\mu_D)$	$\mu_D = 0$ i.e., $\mu_1 = \mu_2$	1. $\mu_D > 0$ 2. $\mu_D < 0$ 3. $\mu_D \neq 0$	$\frac{\bar{d}\sqrt{n}}{s_d}$ where $\bar{d} = \Sigma d_i / n$ and $s_d = \sqrt{\frac{1}{n-1}\Sigma(d_i - \bar{d})^2}$	1. greater than $t_{n-1,\alpha}$ 2. less than $-t_{n-1,\alpha}$ 3. less than $-t_{n-1,\alpha/2}$ or greater than $t_{n-1,\alpha/2}$
6.	Quantitative data; μ_1, σ_1^2, μ_2 and σ_2^2 are unknown; Both populations are normally distributed.	$\sigma_1^2 - \sigma_2^2$	$\sigma_1^2 = \sigma_2^2$	1. $\sigma_1^2 > \sigma_2^2$ 2. $\sigma_1^2 < \sigma_2^2$ 3. $\sigma_1^2 \neq \sigma_2^2$	$F_0 = \frac{S_1^2}{S_2^2}$	$F_0 > F_{\alpha/2, n_1-1, n_2-1}$ $F_0 < F_{1-\alpha/2, n_1-1, n_2-1}$ $F_0 > F_{\alpha, n_1-1, n_2-1}$
7.	Qualitative data; binomial case	$p_1 - p_2$	$p_1 = p_2$	1. $p_1 > p_2$ 2. $p_1 < p_2$ 3. $p_1 \neq p_2$	$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ where $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$ Both n_1 and n_2 are assumed large $\hat{p}_1 = \frac{x_1}{n_1}$ $\hat{p}_2 = \frac{x_2}{n_2}$	1. greater than z_α 2. less than $-z_\alpha$ 3. less than $-z_{\alpha/2}$ or greater than $z_{\alpha/2}$ The level of significance approximately 100α per cent

6.15 SUMMARY

In this chapter, we introduced the second topic in inferential statistics: tests of hypotheses. In a test of hypothesis, we test a certain given theory or belief about a population parameter. Based on some sample proportions, we find out whether or not a given claim (or a statement) about a population parameter is true.

We discussed how to make such tests of hypotheses about the population mean, μ and population proportion, p, for a single population. The hypotheses-testing procedures have been extended to the difference between two population means and the difference between two population proportions. Constructing confidence intervals and testing hypotheses about population parameters are referred to as *making inferences*.

PROBLEMS

P6.1 Determine the critical values for a hypothesis test at the 1% significance level ($\alpha = 0.01$) if the test is

(*a*) two-tailed

(*b*) left-tailed

(*c*) right-tailed

P6.2 Determine the critical values for a hypothesis test at the 5% significance level ($\alpha = 0.05$) if the test is

(*a*) two-tailed

(*b*) left-tailed

(*c*) right-tailed

P6.3 Determine the critical values for a hypothesis test at the 0.5% significance level ($\alpha = 0.005$) if the test is

(*a*) two-tailed

(*b*) left-tailed

(*c*) right-tailed

P6.4 Determine the critical values for a hypothesis test at the 0.2% significance level ($\alpha = 0.002$) if the test is

(*a*) two-tailed

(*b*) left-tailed

(*c*) right-tailed

P6.5 The breaking strength of a machine part is required to be at least 75 Pa. Based on previous test data, the standard deviation of breaking strength is 6 Pa. A random sample of 16 specimens is tested and the average value is found to be 73 Pa.

(*a*) state the hypotheses that need to be tested

(*b*) test the hypotheses using $\alpha = 0.05$.

P6.6 The diameter of copper shafts produced by a certain manufacturing company process should have a mean diameter of 0.51 mm. The diameter is known to have a standard deviation of 0.0002 mm. A random sample of 40 shafts has an average diameter of 0.509 mm. Test the hypotheses on the mean diameter using $\alpha = 0.05$.

P6.7 Refer to Problem P6.7. Determine the sample size required to construct a 95% confidence interval on the mean that has a total width of 2.0.

P6.8 The mean life span of population of light bulbs manufactured by a particular company is found to be 2400 hours with a standard deviation of 250 hours. A sample of 100 light bulbs produced in a lot is found to have a mean life span of 2300 hours. Test whether the difference between the population and sample means is statistically significant. Assume $\alpha = 0.05$.

P6.9 (*a*) In an experimental lab, 10 students are selected at random from a class and were assigned to set up an experiment. Their time to complete the experimental set up are recorded (in minutes) is 51, 55, 54, 50, 56, 59, 57, 59, 58, 60. Would it be appropriate to suggest that the mean completion time for the given experimental set up in the student population as 58.2 minutes? Given $\alpha = 0.05$.

(*b*) The mean weight of all basketball players in a particular college is 69.3 kg. A random sample of 20 such players produced a mean weight of 70.0 kg with a standard deviation of 2.1 kg. Assuming that the weights of all college basketball players are normally distributed, test at the 1% significance level if their mean weight is different from.

P6.10 A study claims that students in a particular college spend an average of 18 hours per week on leisure activities. The administration wanted to test this claim. A sample of 10 students selected at random gave their responses as follows (in hours):

14, 16, 19, 22, 23, 25, 26, 33, 38, 41

Assume that the time spend on leisure activities by all students is normally distributed. Using the 5% significance level, can the administration conclude that the claim of the earlier study is true?

P6.11 A manufacturing company claims that the mean production time for their product is not more than 45 minutes. A random sample of 20 products selected from the production line of the company showed that the mean production time for this sample is 50 minutes with a standard deviation of 3 minutes. Using 1% significance level, can we conclude that the company's claim is true?

P6.12 The mean height of 45 students who showed an above-average participation in college sports was 68.1 inches with a standard deviation of 2.42 inches, while 45 other students who showed no interest in such sports participation and a mean height of 67.67 inches with a standard deviation of 2.76 inches.

(*a*) test the hypothesis that the students who participate in sports activities are taller that the other students.

(*b*) by how much the sample size of each of the two groups be increased in order that the observed difference of 0.75 inches in the mean heights be significant at the level of significance (*i*) 0.05 and (*ii*) 0.01.

P6.13 A garage has a line of electronic tune-up systems. Another line has been installed and the manager would like to find out if the output of the new line is greater than that of the old one. Twelve days of data was gathered at random from the old line and 10 days of data are gathered at random from the new line, with $\overline{X}_1 = 1{,}150$ cases and $\overline{X}_2 = 1{,}175$ cases. It is known that $\sigma_1^2 = 50$ and $\sigma_2^2 = 60$. Test the appropriate hypotheses at $\alpha = 0.05$ assuming that the outputs are normally distributed.

P6.14 A company tested two different machines in order to buy the best one out of the two machines. Machine 1 was run for 8 hrs and it produced an average of 120 parts per hour with a standard deviation of 9 parts. Machine 2 was run for 10 hrs and it produced an average of 111 parts per hour with a standard deviation of 6 parts. The production per hour of both machines is assumed as normally distributed. The standard deviation of the hourly productions of the two populations are unequal. Using a 2.5% significance level, can we conclude that the mean number of parts produced per hour by machine 1 is higher than that of machine 2?

P6.15 To compare two brands of machine parts, Brand *A* and Brand *B*, for their carbon content, a sample of 60 was inspected from Brand *A* and a sample of 40 from Brand *B*. The results of the inspection are given below:

Brand A	$\bar{x} = 16$	$s_1^2 = 3$
Brand B	$\bar{y} = 17.4$	$s_2^2 = 4$

At the 5% significance level, do the two brands differ in their mean carbon content?

P6.16 The time required to change flat tires in a garage is a normal distributed random variable. The tire changing time for 15 automobiles selected at random are as follows:

5 9 7 6 24 11 4 13 10 9 20 8 19 17 25

(*a*) at $\alpha = 0.05$, does it seem reasonable that the mean tire changing time is 10 hours?

(*b*) does it seem reasonable that the mean tire changing time is less than 10 hours?

(*c*) does it seem reasonable that the mean tire changing time is greater than 10 hours?

P6.17 The results of tensile tests on parts made of steel specimens are as follows. The load at which specimen failure in MPa are: the sample mean $\bar{X} = 12$, and the sample standard deviation is $s = 3$. The numbers of samples are 25. Do the data suggest that the mean load at failure exceeds 10 MPa? Assume that the load at failure has normal distribution, and uses $\alpha = 0.05$.

P6.18 The maximum baking temperature for a certain component is supposed to be 90°C. The oven temperature have a standard deviation of 10°C at the maximum baking temperature. A random sample of 228 readings is taken and the average baking temperature is found to be 94°C. Do the data suggest that the mean temperature at baking process exceeds 90°C? Assume that the baking process has a normal distribution and use $\alpha = 0.05$.

P6.19 A company sent six of its sales employees to attend a special course in order to improve the sales of the company's product. The one-week sales of these salespersons before and after attended the course are given below:

Before	9	12	14	16	18	25
After	14	18	19	20	24	24

Using 1% significance level, can we conclude that the mean weekly sales for all salespersons increase as a result of attending the course? Assume the population of paired differences has a standard normal distribution.

P6.20 Table P6.20 gives the systolic blood pressure of seven adults before and after the completion of a special dietary plan for 3 months.

Table P6.20

Before	210	180	195	220	231	199	224
After	192	185	185	223	220	185	230

Consider μ_d as the mean difference between systolic blood pressure before and after completing this special dietary plan of all adults. Using the 5% significance level, can we conclude that the mean of the paired difference μ_d is different from zero? Assume that the population of paired difference is approximately normally distributed.

P6.21 A random sample of 6 automobiles was selected and these automobiles were driven for one week without the gasoline additive and then for one week with the gasoline additive. The manufacturer of the gasoline additive claims that the use of this additive increases gasoline mileage. The following table gives miles per gallon for these automobiles without and with the gasoline additive.

Without	25	28	19	23	15	29
With	27	29	21	24	17	31

Using the 2.5% significance level, can you conclude that the use of the gasoline additive increases the gasoline mileage? Assume that the population of paired differences is approximately normally distributed.

P6.22 A manufacturing company was testing two machines for production automation purpose. The first machine was run for 8 hours and it produced an average of 136 items per hour with a standard deviation of 9 items. The second machine was run for 10 hours and produced an average of 127 items per hour with a standard deviation of 6 items. Assume that the production rates of both machines are normally distributed.

(a) make a 95% confidence interval for the difference between the population means

(b) using a 2.5% significance level, can it be concluded that the mean number of items produced per hour by the first machine is higher than that of the second machine?

P6.23 Refer to problem 6.9. Assume that σ_1^2 and σ_2^2 are unknown and not assumed to be equal. However $S_1^2 = 65$ and $S_2^2 = 75$. Test the appropriate hypotheses at $\alpha = 0.05$.

P6.24 Refer to Problem 6.3. Assume that σ_1^2 and σ_2^2 are unknown but assumed to be equal with sample variances as given. Test the appropriate hypotheses at $\alpha = 0.05$.

P6.25 The annual salaries, in thousands of dollars, of 10 men in middle management at a given corporation are: 55, 65, 68, 70, 52, 56, 60, 72, 71, 73, while those for 8 women are: 56, 49, 58, 51, 61, 55, 57, 59. Let X and Y denote the salaries of men and women, respectively; assuming normal distribution and equal standard deviations, test H_0: $\mu_x = \mu_y$ against H_0: $\mu_x > \mu_y$ at the 0.05 significance level by constructing a critical region for the tests.

P6.26 The following information was obtained from two independent samples selected from two normally distributed populations with unknown but equal standard deviation.

Sample 1	5	8	9	9	10	10	11	12	12	13	13	
Sample 2	11	12	13	14	14	16	16	17	17	18	19	19

Test at the 2.5% significance level if μ_1 is lower than μ_2, where μ_1 and μ_2 are the means of populations 1 and 2 respectively.

P6.27 A consulting agency was asked by a large health insurance company to investigate if business majors were better salespersons. A sample of 40 salespersons with a business degree showed that they sold an average of 10 insurance policies per week with a standard deviation of 1.80 policies. Another sample of 45 salespersons with a degree other than business showed that they sold an average of 8.5 insurance policies per week with a standard deviation of 1.35 policies. At 1% significance level, can we conclude that the persons with a business degree are better salespersons than those who has a non-business degree?

P6.28 The administration wants to test if the mean GPAs (grade point averages) of all male and female college students who actively participate in sports are different. A random sample of 28 male and 24 female students who actively participate in sports was selected. The mean GPAs of the two groups are found to be 2.62 and 2.74 out of 4.0 respectively, with the corresponding standard deviation equal to 0.43 and 0.38. Assume that the GPAs of all male and female students have a normal distribution with the same standard deviation. Test at the 5% significance level if the mean GPAs of the two populations are different.

P6.29 The percentage of manganese in an alloy used in metal castings is measured in 50 randomly selected parts. The sample standard deviation is $s = 0.4$. Test the hypothesis H_0: $\sigma = 0.25$ versus H_1: $\sigma \neq 0.25$ using $\alpha = 0.05$.

P6.30 The thickness of a machine part used in a system is its critical dimension and that measurements of the thickness of a random sample of 18 such parts have a variance of $s^2 = 0.6$, where the measurements are in thousandths of an inch. The process is considered good if the variation of the thickness is given by a variance not greater than 0.3. Assuming that the measurements contribute a random sample from a normal population, test the null hypothesis $\sigma^2 = 0.3$ against the alternative hypotheses $\sigma^2 > 0.3$ at the 0.05 level of significance.

P6.31 In a random sample, the weights of 24 machined parts of a certain process have a standard deviation of 3 grams. Assume that the weights constitute a random sample from a normal population, test the null hypothesis $\sigma = 4$ grams against the two-sided alternative $\sigma \neq 4$ grams at the 0.01 level of significance.

P6.32 The variance of scores on a standardised engineering test for all seniors was 150 in 2003. A sample of scores for 20 seniors who took this test in 2007 gave a variance of 170. Test at the 5% significance level if the variance of current seniors on this test is different from that of 2003. Assume that the scores of all seniors on this test are approximately normally distributed.

P6.33 The gaskets manufactured by a particular company must have a variance of 0.003 mm for acceptance by the customer. A random sample of 29 such gaskets gave a variance of 0.0058 mm. Test at the 1% significance level if the variance of all such gaskets is greater than 0.003 mm. Assume that the thickness of all such gaskets produced by this company are approximately normally distributed.

P6.34 The manufacturer of a certain brand of fluorescent lights claims that the variance of the lives of these is 4000 hours. A consumer agency took a random sample of 25 such lights and tested them. The variance of the lives of these lights was found to be 4800 hours. Assume that the lives of all such lights are approximately normally distributed. Test at the 5% significance level if the variation of such light is different from 4000 hours.

P6.35 A random sample of 15 light bulbs produced by a certain company was selected and each bulb was tested for the number of hours before it burned out. If $s = 13$ hours at the 5% level of significance, is the population standard deviation different from 10 hours?

P6.36 A manufacturing company has two production lines that turn parts to a predetermined length. From a sample of 21 items taken from production line 1 and 30 items taken from production line 2, it was found that $s_1 = 15$ and $s_2 = 12$. Test the hypothesis that the two production lines have the same variance from the predetermined length. Take $\alpha = 0.05$.

P6.37 A new voltage-regulating device is installed in an electromechanical system. Before its installation, a random sample yielded the following information about the percentage of breakdown of the system: $\bar{X}_1 = 8$, $s_1^2 = 36$ and $n_1 = 8$. After installation, a random sample yielded $\bar{X}_2 = 7$, $s_2^2 = 25$ and $n_2 = 9$.

(*a*) can you conclude that the variance are equal? Use $\alpha = 0.05$

(*b*) has the new voltage-regulating device reduced the machine system breakdowns significantly? Use $\alpha = 0.05$.

P6.38 In bolted joints using gaskets the variability in the thickness of the gasket material is a critical characteristic of the bolted joint and a low variability is desirable for better performance of the bolted joint. Two different materials are being considered to determine whether one is superior in reducing the variability of the thickness. Twenty gaskets are selected. The sample standard deviation of gasket thickness are $s_1 = 2.96$ mm and $s_2 = 3.05$ mm, respectively. Is there any evidence to indicate that either gasket material is preferable? Use $\alpha = 0.05$.

P6.39 Two independent random samples are drawn from two normal populations. Test $H_0 : \sigma_1^2 = \sigma_2^2$ at 5% level of significance.

Sample 1	10	8	18	12	9	14	16	9				
Sample 2	8	13	21	15	11	15	18	9	21	22	11	16

P6.40 A random sample of six values of a random variable X is 24, 25, 28, 31, 32, 40 and independently obtained eight sample values of a random value Y are 12, 14, 15, 16, 16, 17, 17, 19. Assume that X and Y are normally distributed, test the null hypothesis $H_0 : \sigma_X^2 = \sigma_Y^2$ against the alternative hypothesis $H_1 : \sigma_X^2 > \sigma_Y^2$ at the 0.05 significance level by finding the critical region for the alternative hypothesis.

P6.41 Refer to Problem P6.40. Check if $H_0 : \sigma_x^2 = \sigma_y^2$ be rejected in favour of $H_1 : \sigma_x^2 > \sigma_y^2$ at the 0.01 significance level.

P6.42 Two different manufacturing processes are used by a company to produce turbine blades with identical mean surface roughness. It is desirable to select the process having the least variability in surface roughness. A random sample of $n_1 = 11$ parts from the first process results in a sample standard deviation $s_1 = 5.1$ micro centimeters, and a random sample of $n_2 = 16$ parts from the second process results in a sample standard deviation of $s_2 = 4.7$ micro centimeters. Find at 90% confidence interval on the ratio of the two standard deviations, s_1/s_2.

P6.43 A manufacturing firm promises a distributor that not more than 0.10 of those delivered toys will be blemished. A random sample of 500 toys delivered over the past 6 months yields 75 blemished toys. Should the toy manufacturer be declared in default at $\alpha = 0.05$?

P6.44 In a survey of 1,000 random samples of machined parts, 700 were found to be non-defective. The manufacturing company wants at least 60% of all parts it produces are non-defective. Using a 0.05 significance level, can we conclude that the company's manufacturing policy is maintained?

P6.45 A political analyst knows that a certain district voted 70% Republican in the last Congressional election. He also knows, however, that drastic changes in this district have occurred since the last election. He believes that these changes may have changed the percentage of votes that will vote for the Republican candidate for Congress, but doesn't know which way it will be changed. He polls 2,000 voters in the district and finds that 960 of these plan to vote for the Democratic candidate. Is his belief supported, at 0.01 level of significance?

P6.46 A sample survey shows that 615 of 820 people surveyed prefer to live in single homes. Test the hypothesis that the true proportion of people who prefer to live in single homes is 0.70. Given $\alpha = 0.05$.

P6.47 A machine is found to produce 32 per cent defective bolts. After repairing the machine, it was found it produced 24 defective bolts in the first production of 100. Is the true proportion of defective tubes reduced after the repairs? Use $\alpha = 0.01$.

P6.48 In a certain city only 400 men out of 900 were found to drink alcohol. Determine whether this information supports the view that majority of men in that city are non-alcohol drinkers. (Use 5% level of significance).

P6.49 It was found that the proportion of families in the U.S. who took vacation of at least 1 week last year was 20 per cent. In order to find the attitude of families on traveling this year, 100 families selected at random and were interviewed and, of these, 15 said they would take such a vacation. Has the attitude changed from last year? Use $\alpha = 0.10$.

P6.50 Two different polishing materials are being evaluated for use in machining operations in a manufacturing company. One hundred machine parts were polished using polishing material *A* and of this number 87 had no polishing-included defects. Another 100 machine parts were polished using polishing material *B* and 75 parts were satisfactory upon completion. Is there any reason to believe that the two polishing materials differ? Use $\alpha = 0.01$.

P6.51 Two different types of grinding machines are used to finish metal parts. A part is considered defective if it has excessive surface roughness or variation. Two random samples, each of size 100 are selected and 5 defective parts are found in the sample from machine 1, while 3 defective parts are found in the sample from machine 2. Is it reasonable to conclude that both machines produce the same fraction of defective parts, using $\alpha = 0.05$?

P6.52 A sample of 500 observations taken from the first population gave $\bar{X}_1 = 400$. Another sample of 600 observations taken from the second population gave $\bar{X}_2 = 420$.

(a) find the point estimate of $\hat{p}_1 - \hat{p}_2$

(b) make a 97% confidence interval for $\hat{p}_1 - \hat{p}_2$

(c) show the rejection and non-rejection regions on the sampling distribution of $\hat{p}_1 - \hat{p}_2$ for $H_0: p_1 = p_2$ versus $H_1: p_1 > p_2$. Use a significance level of 2.5%

(d) find the value of the test statistic z for the test of part *(c)*

(e) do you reject the null hypothesis mentioned in part *(c)* at a significance level of 2.5%?

P6.53 According to a particular research study, 24% of children in the age group 3 to 17 were overweight in 1994 and 31% were considered overweight in 2003. These percentages are based on random samples of 400 and 500 children in the given age group in 1994 and 2003, respectively. Test at the 1% significance level if the proportion of overweight children in the age group was less than in 1994 than in 2003.

P6.54 A sample survey conducted in a city results show that out of 800 men 480 are employed whereas out of 600 women only 350 were employed. Can this difference between two proportions of employed persons be ascribed due to sampling fluctuations? Use $\alpha = 0.05$.

P6.55 An insecticide of Brand *A* was sprayed into a jar containing 12 insects and it was found 95 of the insects were killed. When another jar containing 145 insecticides of the same type was sprayed with Brand *B*, 124 insects were killed. Do the two brands differ in their effectiveness at 2% level of significance?

P6.56 A study polled independent random samples of 747 men and 434 women in a particular city. Of those sampled, 276 men and 195 women responded that they frequently ordered pure vegetarian meals. Do the data poll provide sufficient evidence to conclude that the percentage of men who sometimes order pure vegetarian meals is smaller than the percentage of women who sometimes order vegetarian meals? Use 5% significance level.

REVIEW QUESTIONS

1. Describe the meaning of each of the following terms:

   (a) null hypothesis
   (b) alternate hypothesis
   (c) critical points
   (d) significance level
   (e) rejection region
   (f) nonrejection region
   (g) tails of a test
   (h) two types of errors

2. Describe briefly Type I and Type II errors.
3. Describe the general steps involved in testing a hypothesis.
4. Explain the difference between the critical value of z and observed value of z.
5. What does the level of significance represent in a test of hypothesis?
6. Describe the conditions that must hold true to use in t-distribution to make a test of hypothesis about the population mean.
7. Explain when a sample is large enough to use the normal distribution to make a test of hypothesis about the population proportion.
8. Describe the following terms:

   (a) α
   (b) β
   (c) left-tailed test
   (d) right-tailed test
   (e) one-tailed test
   (f) test statistic

9. Explain the meaning of independent and dependent samples.
10. Describe the sampling distribution of $\bar{x}_1 - \bar{x}_2$ for large and independent samples. What are the mean and standard deviation of this sampling distribution?
11. Describe the conditions that must hold true to use the t-distribution to make a confidence interval and to test a hypothesis about $\mu_1 - \mu_2$ for two independent samples selected from two populations with unknown but equal standard deviations.

12. Explain when the paired sample procedures are used to make confidence intervals and test hypothesis.
13. Describe the shape of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ for two large samples. What are the mean and standard deviation of this sampling distributions?
14. Define the following terms:

    (a) paired or matched samples

    (b) μ_d

    (c) s_d

    (d) σ_d

STATE TRUE OR FALSE

1. Null hypothesis is a hypothesis to be tested. (True/False)
2. Alternative hypothesis is a hypothesis to be considered as an alternative to the null hypothesis. (True/False)
3. The problems in a hypothesis test is to decide whether the null hypothesis should be accepted in favour of the alternative hypothesis. (True/False)
4. A hypothesis test whose alternative hypothesis has the form H_1: $\mu \neq \mu_0$ is called a one-tailed test. (True/False)
5. A hypothesis test whose alternative hypothesis has the form H_1: $\mu < \mu_0$ is called a right-tailed test. (True/False)
6. A hypothesis test whose alternative hypothesis has the form H_1: $\mu > \mu_0$ is called a right-tailed test. (True/False)
7. A hypothesis is a statement that something is true. (True/False)
8. The decision criterion provides an objective method for deciding whether the null hypothesis should be rejected in favour of the alternative hypothesis. (True/False)
9. Test statistic is used as a basis for deciding whether the null hypothesis should be rejected. (True/False)
10. Rejection region is the set of values for the test statistic that leads to acceptance of the null hypothesis. (True/False)
11. Nonrejection region is the set of values for the test statistic that leads to non rejection of the null hypothesis. (True/False)
12. Critical values are the values of the test statistic that separate the rejection and nonrejection regions. (True/False)
13. A critical value is considered part of the nonrejection region. (True/False)
14. Type I error is accepting the null hypothesis when it is in fact true. (True/False)
15. Type II error is not rejecting the null hypothesis when it is in fact false. (True/False)
16. The probability of making a Type I error, that is, of rejecting a true null hypothesis, is called the significance level, α, of a hypothesis test. (True/False)
17. For a fixed-sample size, the smaller we specify the significance level, α, the large will be the probability, β, of not rejecting a false null hypothesis. (True/False)
18. If a null hypothesis is not rejected, we conclude that the alternative hypothesis is true. (True/False)

19. If the null hypothesis is not rejected, we conclude that the data do not provide sufficient evidence to support the alternative hypothesis. (True/False)
20. If it is important not to reject to a true null hypothesis, the hypothesis test should be performed at a small significance level. (True/False)
21. For a fixed sample size, decreasing the significance level of a hypothesis test results in an increase in the probability of making a Type II error. (True/False)
22. Hypothesis tests and confidence intervals are not closely related. (True/False)
23. The power of a hypothesis test is the probability of not making a Type II error, that is, the probability of rejecting a false null hypothesis. (True/False)
24. Type I error is rejecting a true null hypothesis. (True/False)
25. Type II error not rejecting a true null hypothesis. (True/False)
26. Significance level is the probability of rejecting a true null hypothesis, that is, of making a Type I error. (True/False)
27. In the context of hypothesis testing, α is the probability of making a Type II error (rejecting a true null hypothesis) also known as the significance level of the hypothesis test. (True/False)
28. In the context of hypothesis testing β is the probability of making a Type II error (not rejecting a false null hypothesis). (True/False)
29. In the context of hypothesis testing, $(1 - \beta)$ is the power of the hypothesis test (the probability of not making a Type II error or equivalently) of rejecting a false null hypothesis. (True/False)
30. For a fixed significance level, increasing the sample size decreases the power. (True/False)
31. Decreasing the significance level of a hypothesis test without changing the sample size decreases the probability of a Type II error, equivalently, decreases the power. (True/False)
32. The *P*-value of a hypothesis test equals the smallest significance level at which the null hypothesis can be rejected, that is, the smallest significance level for which the observed sample data results in rejecting of H_0. (True/False)
33. If the *P*-value is less than or equal to the specified significance level, reject the null hypothesis, otherwise, do not reject the null hypothesis. (True/False)
34. Independent sample means that the sample selected from one of the populations has no effect or bearing on the sample selected from the other population. (True/False)
35. Population proportion is the proportion (percentage) of the entire population that has the specified attribute. (True/False)
36. Sample proportion is the proportion (percentage) of a sample from the population that has the specified attribute. (True/False)
37. A population proportion is a parameter because it is a descriptive measure for a population. (True/False)
38. A sample proportion is a statistic because it is a descriptive measure for a sample. (True/False)
39. Number of successes is the number of members in the sample that have the specified attribute. (True/False)
40. Number of failures is the number of members in the sample that have the specified attribute. (True/False)

41. The mean of all possible sample proportions is equal to the population proportion. (True/False)
42. For large sample, the possible sample proportions have approximately a chi-square distribution. (True/False)
43. A rule of thumb for using a normal distribution to approximate the distribution of all possible sample proportions is that both np, $n(1-p)$ are 5 or greater. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **True** 3. **False** 4. **False** 5. **False** 6. **True** 7. **True** 8. **True** 9. **True** 10. **False**
11. **True** 12. **True** 13. **False** 14. **False** 15. **True** 16. **True** 17. **True** 18. **True** 19. **True** 20. **True**
21. **True** 22. **False** 23. **True** 24. **True** 25. **False** 26. **True** 27. **True** 28. **True** 29. **True** 30. **False**
31. **False** 32. **True** 33. **True** 34. **True** 35. **True** 36. **True** 37. **True** 38. **True** 39. **True** 40. **False**
41. **True** 42. **False** 43. **True**

❑❑❑

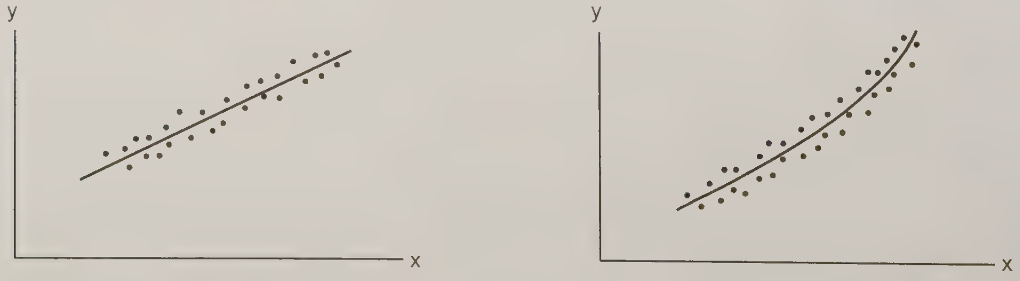
CHAPTER 7

Curve Fitting, Regression and Correlation

In real life engineering practice, often a relationship is found to exist between two (or more) variables. For example: the experimental data for force (N) and velocity (m/s) from a wind tunnel experiment. A mechanical element/component is suspended in a wind tunnel and the force measured for various levels of wind velocity. This relationship can be visualised by plotting force versus velocity. It is frequently desirable to express this relationship in mathematical/analytical form by establishing an equation connecting the variables.

In order to determine an equation connecting the variables, it is often necessary to collect the data depicting the values of the variables under consideration.

For example, if x and y denote respectively the velocity and force from the wind tunnel experiment, then a sample of n individual would give the velocities $x_1, x_2, \ldots, x_n$ and the corresponding forces $y_1, y_2, \ldots, y_n$. When these points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ are plotted on a rectangular coordinate system, the resulting set of points on the plot is called the *scatter diagram*. From such a scatter diagram, one can visualise a smooth curve approximating the given data points. Such a curve is known as an *approximating curve*. Figure 7.1(a) shows that the data appears to be approximated by a straight line and it clearly exhibits a *linear relationship* between the two variables. On the other hand Fig.7.1(b) shows a relationship which is not linear and in fact it is a nonlinear relationship between the variables. *Curve fitting* is the general problem of finding equations of approximating curves which best fit the given set of data.



(a) Linear Relationship

(b) Non-linear Relationship

Fig. 7.1: Linear and non-linear relationship

Approximating Curves

Table 7.1 lists a few common approximating curves used in practice and their equations. The variables x and y are called the *independent* and *dependent variables* respectively. The scatter diagrams of the variables or transformed variables will help determine the curve to be used.

Table 7.1

No.	Equation	Description of the curve
1.	$y = a + bx$	Straight line
2.	$y = a + bx + cx^2$	Parabola or quadratic curve
3.	$y = a + bx + cx^2 + dx^3$	Cubic curve
4.	$y = a + bx + cx^2 + dx^3 + ex^4$	Quadratic curve
5.	$y = a_0 + a_1x + a_2x^2 + \ldots + a_nx^n$	n^{th} degree curve
6.	$y = \dfrac{1}{c+mx}$ or $\dfrac{1}{y} = c + mx$	Hyperbola
7.	$y = ab^x$ or $\log y = \log a + x(\log b) = a_0 + b_0x$	Exponential curve
8.	$y = ax^b$ *or* $\log y = \log a + b(\log x)$	Geometric curve (power function)
9.	$y = ab^x + c$	Modified exponential curve
10	$y = ax^b + c$	Modified geometric curve
11.	$y = pq^{b^x}$ or $\log y = \log p + b^x \log q = ab^x + q$	Gompertz curve
12.	$y = pq^{b^x} + h$	Modified Gompertz curve
13.	$y = \dfrac{1}{ab^x + q}$ or $\dfrac{1}{y} = ab^x + q$	Logistic curve
14.	$y = be^{mx}$ or $y = b\,10^{mx}$	Exponential function
15.	$y = \dfrac{1}{mx+b}$	Reciprocal function
16.	$y = \alpha\dfrac{x}{\beta + x}$	Saturation-growth-rate equation

In Table 7.1, $a, b, c, d, e, a_0, a_1, a_2, \ldots, a_n, b_0, p, q, h, m, a$ and b are all constant coefficients.

Linear Regression

Linear *regression* and *correlation* are two commonly used methods for examining the relationship between quantitative variables and for making predictions. In this chapter, we review linear equations with one independent variable, explain how to find the *regression equation*, the equation of the line that best fits a set of data points. We also examine the coefficient of determination that is a descriptive measure of the utility of the regression equation for making predictions. In addition, we discuss the linear correlation coefficient, which provides a descriptive measure of the strength of the linear relationship between the two quantitative variables.

The inferential methods in regression and correlation are also covered in this chapter.

7.1 LINEAR EQUATION

The general form of a *linear equation* with one independent variable can be written as

$$y = a + bx$$

where a and b are constants (fixed numbers), x is the independent variable, and y is the dependent variable. The graph of a linear equation with one independent variable is a *straight line*, or simply a *line*. Also, any non vertical line can be represented by such an equation.

Linear equations with one independent variable occur frequently in applications of mathematics to many different fields, including the social sciences, engineering, and management as well as physical and mathematical sciences.

For a linear equation $y = a + bx$, the number a is the y-value of the point of intersection of the line and the y-axis. The number b measures the steepness of the line. b indicates how much the y-value changes when the x-value increases by 1 unit. Figure 7.2(a) illustrates these relationships.

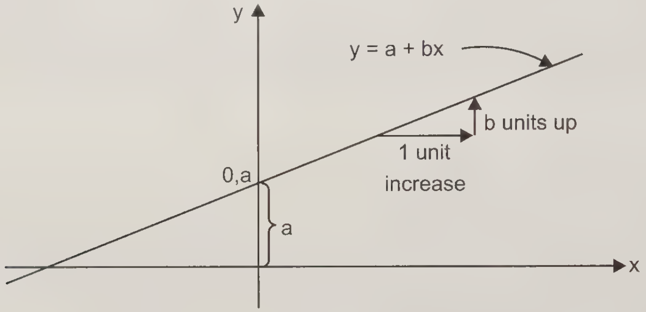


Fig. 7.2(a): Graph of y = a + bx

The numbers a and b have special names that reflect these geometric interpretations. For a linear equation $y = a + bx$, the number a is called the *y-intercept* and the number b is called the *slope*.

The graph of the linear equation $y = a + bx$ slopes upward if $b > 0$, slopes downward if $b < 0$, and is horizontal if $b = 0$, as shown in Fig. 7.2(b).

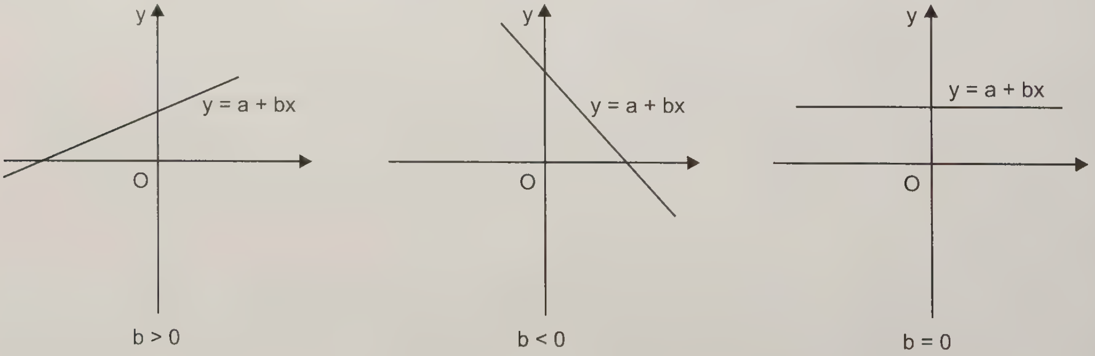


Fig. 7.2(b): Graphical interpretation of slope

7.2 CURVE FITTING WITH A LINEAR EQUATION

Curve fitting is a procedure in which a mathematical formula (equation) is used to best fit a given set of data points. The objective is to find a function that fits the data overall. Curve fitting is used when the values of the data points have some error, or scatter and require a curve fit to the points. Curve fitting can be accomplished with many types of functions and with polynomials of various orders.

Curve fitting using a linear equation (first degree polynomial) is the process by which an equation of the form

$$y = a + bx \tag{7.1}$$

is used to best fit the given data points. This can be accomplished by finding the constants a and b that give the smallest error when the data points are substituted in Eq. (7.1). If the data points consists of only two points, the constants can be obtained such that Eq. (7.1) gives the exact values at the points. Figure 7.3 shows the straight line corresponding to the Eq. (7.1) and passing through the two points. When the data has more than two points, the constants a and b are determined such that the line has the best fit overall as shown in Fig. 7.4.

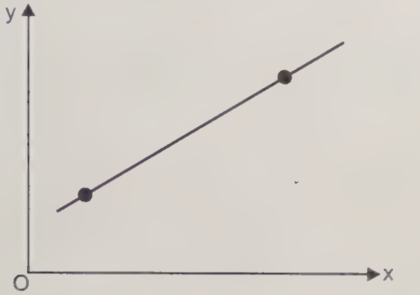


Fig. 7.3: Straight line connecting two points

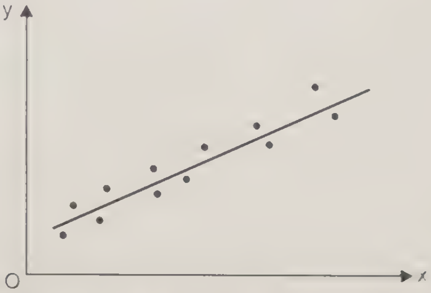


Fig. 7.4: A straight line passing through many data points

The procedure for obtaining the constants a and b that give the best fit requires a definition of best fit and an analytical procedure for deriving the constants a and b. The fitting between the given data points and an approximating linear function is obtained by first computing the error, also called the *residual*, which is the difference between a data point and the value of the approximating function, at each point. Figure 7.5 shows a linear function (straight line) that is used for curve fitting n points.

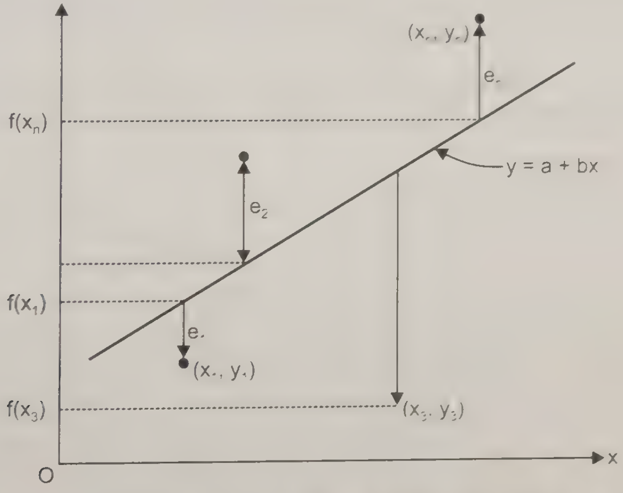


Fig. 7.5: Curve-fitting points with a linear equation y = a + bx

Thus, the *residual*, e is the discrepancy between the true value of y and the approximating value, $a + bx$, predicted by the linear equation.

7.3 CRITERIA FOR A "BEST" FIT

A criterion that measures how well the approximating function fits the given data can be determined by computing a total error E in terms of the residuals as

$$E = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} [y_i - (a + bx_i)] \tag{7.2}$$

where n = total number of points.

However, this is an inadequate criterion, as illustrated in Fig.7.6, which shows that E is zero since $e_1 = -e_4$ and $e_2 = -e_3$.

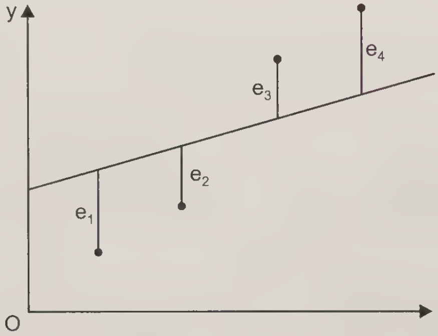


Fig. 7.6: Straight line fit with E = 0

One way to remove the effect of the signs might be to minimize the sum of the absolute values of the discrepancies:

$$E = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} | y_i - a - bx_i | \tag{7.3}$$

Figure 7.7 shows why this criterion is also inadequate. For four points show, for the same set of points there can be several functions that give the same total error. E is the same for the two approximating lines in Figure 7.7.

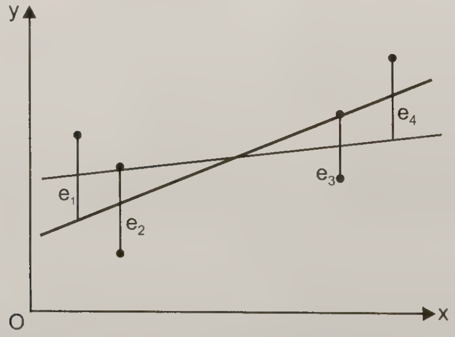


Fig. 7.7: Two straight line fits with the same total error as per Eq. (7.3)

A third strategy for fitting a best line is the *minmax* criterion. In this technique, the straight line is chosen that minimizes the maximum distance that an individual point falls from the line. Again as shown in Fig. 7.8, this technique gives undue influence to an outlier (a single point with a large error).

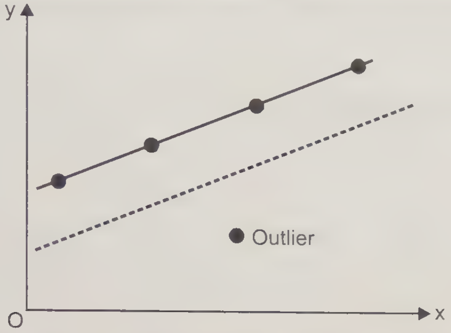


Fig. 7.8: Minimization of the maximum error of any individual point

A method that overcomes the shortcomings of the aforementioned approaches is to minimize the sum of the squares of the residuals:

$$S_r = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{7.4}$$

Equation (7.4) always gives a positive number of S_r (total error) and positive and negative residuals do not cancel each other. This criterion (Eq. 7.4) is called the *least squares* and has many advantages, including that it gives a unique line for a given set of data.

Equation (7.4) can be used to determine the coefficients a and b in the linear function $y = a + bx$ that yield the smallest total error. This is accomplished by using a procedure called linear least-squares regression, which is presented in the next section.

7.4 LINEAR LEAST-SQUARES REGRESSION

Linear least-squares regression is a method in which the coefficients a and b of a linear function $y = a + bx$ are determined such that the function has the best fit to a given set of data points. The best fit is defined as the smallest possible total error that is computed by adding the squares of the residuals according to Eq. (7.4).

For a given set of n data points (x_i, y_i), the overall error calculated by Eq. (7.4) is

$$S_r = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \tag{7.5}$$

Since the values of x_i and y_i are known, S_r in Eq. (7.5) is a non-linear function of two variables a and b. This function S_r has a minimum at the values of a and b where the partial derivatives of S_r with respect to each variable is equal to zero.

Taking the partial derivatives and setting then equal to zero gives

$$\frac{\partial S_r}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \tag{7.6}$$

$$\frac{\partial S_r}{\partial b} = -2\sum_{i=1}^{n}[(y_i - a - bx_i)x_i] = 0 \tag{7.7}$$

Equations (7.6) and (7.7) are a system to two linear equations for the two unknowns *a* and *b*, and can be rewritten in the form

$$na + \left[\sum_{i=1}^{n} x_i\right] b = \sum_{i=1}^{n} y_i \tag{7.8}$$

$$\left[\sum_{i=1}^{n} x_i\right] a + \left[\sum_{i=1}^{n} x_i^2\right] b = \sum_{i=1}^{n} x_i y_i \tag{7.9}$$

Equations (7.8) and (7.9) are called the *normal equations* and can be solved simultaneously for

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n\sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \tag{7.10}$$

or

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \dfrac{\left[\sum_{i=1}^{n} x_i\right]\left[\sum_{i=1}^{n} y_i\right]}{n}}{\sum_{i=1}^{n} x_i^2 - \dfrac{\left[\sum_{i=1}^{n} x_i\right]^2}{n}} \tag{7.11}$$

and

$$a = \frac{\left[\sum_{i=1}^{n} x_i^2\right]\left[\sum_{i=1}^{n} y_i\right] - \left[\sum_{i=1}^{n} x_i y_i\right]\left[\sum_{i=1}^{n} x_i\right]}{n\sum_{i=1}^{n} x_i^2 - \left[\sum_{i=1}^{n} x_i\right]^2} \tag{7.12}$$

or

$$a = \overline{y} - b\overline{x}$$

Since Eqs. (7.10) to (7.12) contain summations that are the same for a set of *n* points, they can also be written as,

$$SS_{xx} = \Sigma(x_i - \overline{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n \tag{7.13}$$

$$SS_{xy} = \Sigma(x_i - \overline{x})(y_i - \overline{y}) = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i) / n \tag{7.14}$$

$$SS_{yy} = \Sigma(y_i - \overline{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n \tag{7.15}$$

The regression equation for a set of *n* data points is

$$\hat{y} = a + bx$$

where $$b = \frac{SS_{xy}}{SS_{xx}} \tag{7.16}$$

and $$a = \frac{1}{n}(\Sigma y_i - b\Sigma x_i) = \overline{y} - b\overline{x} \tag{7.17}$$

7.5 LINEAR REGRESSION ANALYSIS

A regression model is a mathematical equation that describes the relationship between two or more variables. A *single regression* model includes only two variables: one independent and one dependent. The relationship between two variables in a regression analysis is expressed by a mathematical equation called a *regression equation* or *model*. A regression equation that gives a straight line relationship between two variables is called a *linear regression model*; otherwise, it is called a *nonlinear regression model*. Figure 7.9(a) and (b) show a linear and a nonlinear relationship between independent variable and the dependent variable.

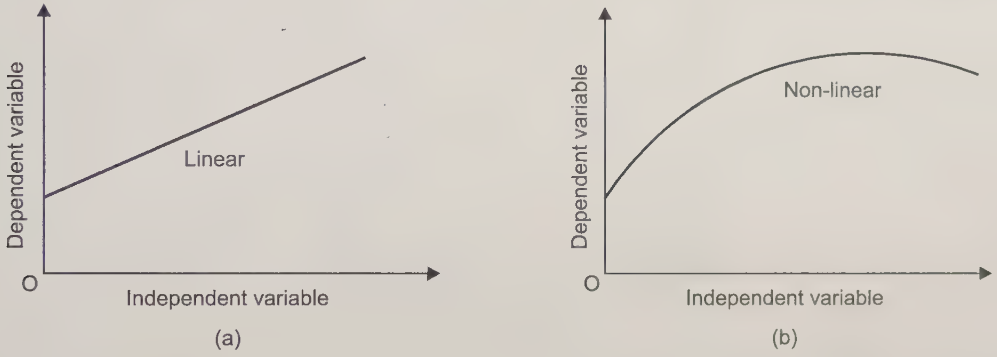


Fig. 7.9: (a) Linear relationship, (b) Non-linear relationship

The equation of a linear relationship between two variables x and y is written as

$$y = a + bx \tag{7.17a}$$

where a gives the y-intercept and b represents the slope of the line.

In regression model, x is the independent variable and y is the dependent variable. The simple linear regression model for population is written as

$$y = A + Bx \tag{7.17b}$$

Equation (7.17b) is called a *deterministic model*. It gives an *exact relationship* between x and y. However, in many instances the relationship between the variables is not exact. Therefore, the complete regression model is then written as

$$y A + Bx + \in \tag{7.17c}$$

where $\in$ is called *the random error term*. This regression model (Eq. (7.5)) is called a *probabilistic model* (or a *statistical relationship*). The random error term $\in$ is included in the model to take into consideration of the following two phenomena:

(a) Missing or omitted variables: The random error term $\in$ is included to capture the effect of all the missing or omitted variables that were not included in the model.

(b) Random variation: The random error term $\in$ is included to capture the random variation.

In Eq. (7.17c), A and B are the *population parameters*. The regression line obtained from Eq. (7.17c) by using the population data is called the *population regression line*. The values of A and B is the population regression line is called the true values of the *y-intercept and slope*.

However, most often the population data is difficult to obtain. As a consequence, we almost always use the sample data and use the model given by Eq. (7.17c). The values of the y-intercept and slope calculated from sample data on x and y are called the *estimated values* of A and B and are denoted by a and b.

The estimated regression model is then written as

$$\hat{y} = a + bx \tag{7.17d}$$

where $\hat{y}$ is the *estimated* or *predicted value* of y for a given value of x. Equation (7.17d) is called the *estimated regression model*. It gives the regression of y on x. A plot of paired observation is called a *scatter diagram* as shown in Fig. 7.10.

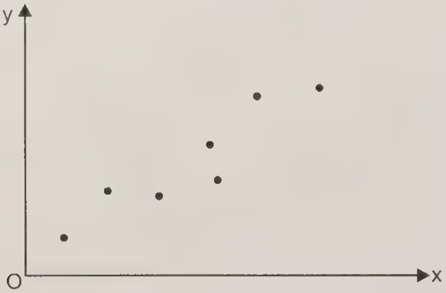


Fig. 7.10: Scatter diagram

To find the line that best fits the scatter of points, we minimize the *error sum of squares*, denoted by SSE, which is given by

$$\text{SSE} = \Sigma e^2 = \Sigma(y - \hat{y})^2 \tag{7.17e}$$

where $e = y - \hat{y}$.

The least-squares method gives the values of a and b such that the sum of squared errors (SSE) is minimum.

For the least-squares regression line $\hat{y} = a + bx$ from Eqs. (7.10) to (7.17), we have

$$b = \frac{SS_{xy}}{SS_{xx}} \text{ and } a = \bar{y} - b\bar{x} \tag{7.18}$$

where $$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

$$SS_{yy} = \Sigma(y - \bar{y}) = \Sigma y^2 - (\Sigma y)^2 / n \tag{7.19}$$

The least-squares regression line = $a + bx$ is also called the *regression of y on x*.

Example E7.1

Table E7.1 gives experimental data for force (N) and velocity (m/s) for an object suspended in a wind tunnel.

Table E7.1

Velocity, v(m/s)	10	20	30	40	50	60	70	80
Force F(N)	24	68	378	552	608	1218	831	1452

(*a*) use the linear least-squares regression to determine the coefficients a and b in the function $\hat{y} = a + bx$ that best fits the data

(*b*) estimate the force when the velocity is 55 m/s.

SOLUTION:

Here $n = 8$.

n	x	y	x^2	xy
1	10	24	100	240
2	20	68	400	1360
3	30	378	900	11340
4	40	552	1600	22080
5	50	608	2500	30400
6	60	1218	3600	73080
7	70	831	4900	58170
8	80	1452	6400	116160
Σ	360	5131	20400	312830

$$\bar{x} = \frac{\Sigma x}{n} = \frac{360}{8} = 45$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{5131}{8} = 641.375$$

From Eq. (7.16), we have

$$b = \frac{SS_{xy}}{SS_{xx}}$$

where $$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 312830 - \frac{(360)(5131)}{8} = 81935$$

and $$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 20400 - \frac{(360)^2}{8} = 4200$$

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{81935}{4200} = 19.5083$$

From Eq. (7.17), we have

$$a = \bar{y} - b\bar{x} = 641.375 - (19.5083)(45) = -236.50$$

Hence $$\hat{y} = -236.50 + 19.5083x$$

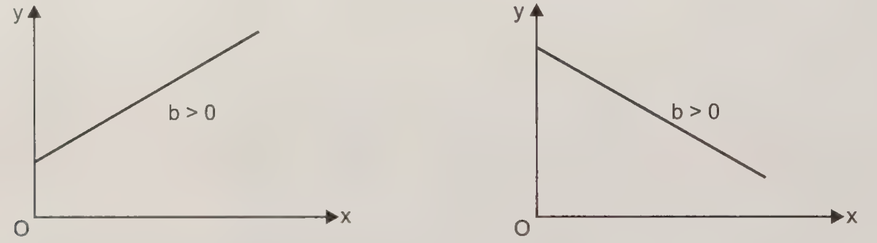
(*b*) The estimated value of the force when the velocity is 55 m/s is given by

$$\hat{y} = a + bx = -236.50 + 19.5083(55) = 836.4583 \text{ N}$$

7.6 INTERPRETATION OF *a* AND *b*

When b is positive, an increment in x will lead to an increase in y and a decrease in x will lead to a decrease in y. That is, when b is positive, the movements in x and y are in the same direction. Such a relationship between x and y is called a *positive linear relationship*. The regression line slopes upward from left to right.

Similarly, if the value of b is negative, an increase in x will cause a decrease in y and a decrease in x will cause an increase in y. The changes in x and y are in opposite directions. Such a relationship between x and y is called a *negative linear relationship*. The regression line slopes downward from left to right. Figure 7.11 shows these two relationships.



(a) Positive linear relationship (b > 0)

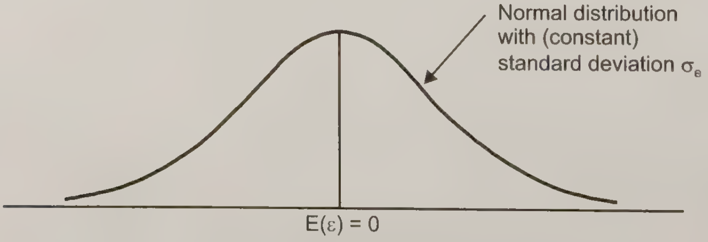
(b) Negative linear relationship (b < 0)

Fig. 7.11: Positive and negative relationship between x and y

Assumptions in the Regression Model

The linear regression analysis is based on the following assumptions:

1. The random error term $\in$ has a mean equal to zero for each x.
2. The errors associated with different observations are independent.
3. For any given x, the distribution of errors is normal.
4. The distribution of population errors for each x has the same (constant) standard deviation, which is denoted by σ_e as shown in Figs. 7.12(*a*) and (*b*).



(a)

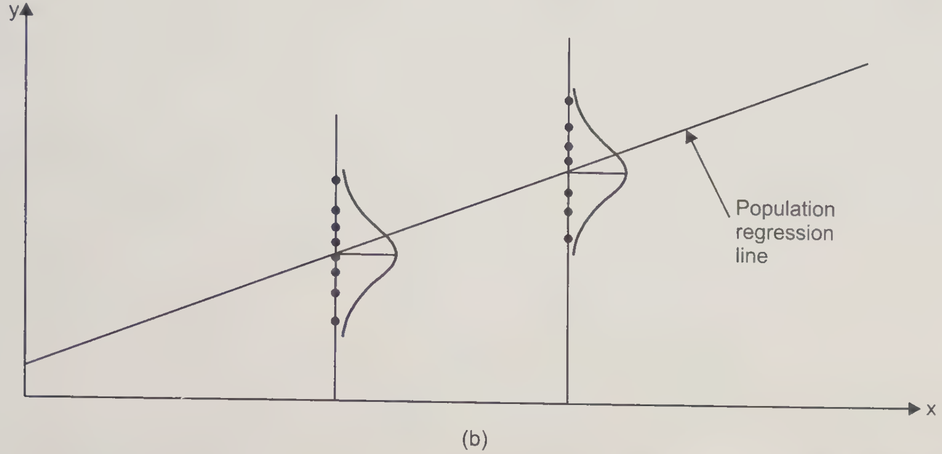


Fig. 7.12: (a) Assumptions in the regression model, (b) Distribution on the regression line

7.7 STANDARD DEVIATION OF RANDOM ERRORS

The standard deviation σ_e measures the spread of the errors around the regression line as shown in Fig. 7.12(b). The *standard deviation of errors* is calculated using

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$$

where $$\text{SSE} = \Sigma(y-\hat{y})^2$$

or $$s_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}} \qquad (7.20)$$

where $$SS_{yy} = \Sigma(y-\bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} \qquad (7.21)$$

In Eq. (7.20), $(n - 2)$ represents the degrees of freedom for the regression model. The reason for $df = n - 2$ is that we lose one degree of freedom to calculate $\bar{x}$ and one for $\bar{y}$.

7.8 COEFFICIENT OF DETERMINATION

The *coefficient of determination*, denoted by r^2, represents the proportion of the *total sum of squares* that is explained by the use of the regression model. The computational formula for r^2 is given by

$$r^2 = b\frac{SS_{xy}}{SS_{yy}}, \qquad 0 \le r^2 \le 1 \qquad (7.22)$$

The *total sum of squares* denoted by SST is the total variation in the observed values of the response variable

$$\text{SST} = SS_{yy} = \Sigma(y-\bar{y})^2 = \Sigma y^2 - \frac{(\Sigma y)^2}{n} \tag{7.23}$$

which is the same formula we use for SS_{yy}.

The regression sum of squares, denoted by SSR, is the variation in the observed values of the response variable explained by the regression:

$$\text{SSE} = \Sigma(y-\hat{y})^2$$

$$\text{SSR} = \text{SST} - \text{SSE}$$

or

$$\text{SST} = \text{SSR} + \text{SSE} \tag{7.24}$$

The ratio of SSR to SST gives the coefficient of determination. That is,

$$r^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SST–SSE}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{7.25}$$

This formula shows that we can also interpret the coefficient of determination as the percentage reduction obtained in the total squared error by using the regression equation instead of the mean, $\bar{y}$, to predict the observed values of the response variable.

The *coefficient of determination,* denoted by r^2, represents the portion of SST that is *explained* by the use of the regression model. Hence,

$$r^2 = \frac{b\, SS_{xy}}{SS_{yy}} \tag{7.26}$$

and

$$0 \le r^2 \le 1$$

The value of r^2 is the proportion of the variation in y that is explained by the linear relationship between x and y. The coefficient of determination is the amount of the variation in y that is explained by the regression line. It is computed as

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}$$

The coefficient of determination, r^2, always lies between 0 and 1. A value of r^2 near 0 suggests that the regression equation is not very useful for making predictions, whereas a value of r^2 near 1 suggests that the regression equation is quite useful for making predictions.

Example E7.2

For the data of Example E7.1, calculate the

(a) standard deviation of errors, s_e

(b) error sum of squares, SSE

(c) total sum of squares, SST

(d) regression sum of squares, SSR

(e) the coefficient of determination, r^2.

SOLUTION:

Referring to Table E7.2, we have

$$n = 8, \Sigma x = 360, \Sigma y = 5131, \Sigma x^2 = 20400, \Sigma xy = 312830, \Sigma y^2 = 5104841$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{360}{8} = 45$$

$$\bar{y} = \frac{\Sigma y}{n} = \frac{5131}{8} = 641.375$$

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 312830 - \frac{(360)(5131)}{8} = 81935$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 20400 - \frac{(360)^2}{8} = 4200$$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 5104841 - \frac{(5131)^2}{8} = 1813945.875$$

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{81935}{4200} = 19.5083$$

$$a = \bar{y} - bx = 641.375 - 19.5083x$$

(a) The standard deviation of errors, s_e

$$s_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}} = \sqrt{\frac{1813945.875 - 19.5083(81935)}{8-2}} = 189.5304$$

(b) The error sum of squares, SSE

$$\text{SSE} = \Sigma(y - \hat{y})^2 = 215530.5833$$

(c) Total sum of squares, SST

$$\text{SST} = SS_{yy} = 1813945.875$$

(d) The regression sum of squares, SSR

$$\text{SSR} = \text{SST} - \text{SSE} = 1813946.875 - 215530.5833 = 1598415.2917$$

(e) The coefficient of determination, r^2

$$r^2 = \frac{b\,SS_{xy}}{SS_{yy}} = \frac{(19.5083)(81935)}{1813945.875} = 0.8812$$

7.9 SAMPLING DISTRIBUTION OF *b*

In most cases, the regression line is estimated using sample data. Inferences are then made about the population regression line based on the sample regression line. The slope *b* of a sample regression line is a point estimator of the slope *B* of the population regression line. Different values of *b* would result in different sample regression lines estimations based on different samples taken from the same population. Therefore, b is a random variable and it possesses a probability distribution that is more commonly called its sampling distribution.

We describe here the shape of the sampling distribution of b, its mean, and standard deviation. The sampling distribution of b is normal based on the assumption of normally distributed random errors.

The mean and standard deviation of b, denoted by σ_b and σ_b, respectively, are

$$\mu_b = B$$

and

$$\sigma_b = \frac{\sigma_e}{\sqrt{SS_{xx}}} \tag{7.27}$$

In general, the standard deviation of population errors σ_e is not known. Therefore, the sample standard deviation of errors σ_e is used to estimate σ_e. In this case, when σ_e is unknown, the standard deviation of b is estimated by σ_b given by

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} \tag{7.28}$$

If the sample size is large ($n \geq 30$) and σ_e is unknown, then the normal distribution can be used to make inferences about B. However, when σ_e is unknown and the sample size is small ($n < 30$), then the normal distribution is replaced by the t-distribution to make inferences about B.

Estimation of B

The value of b computed from the sample regression line is a point estimate of the slope B of population regression line. When σ_e is unknown and the sample size is small ($n < 30$), then the t-distribution is used to make a confidence interval for B.

Confidence Interval for B

The $(1 - \alpha)$ 100% confidence interval for B is given by

$$b \pm t\, s_b \tag{7.29}$$

where

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} \tag{7.30}$$

The value of t is obtained from the t-distribution table (Appendix-G) for $\alpha/2$ area in the right tail of the t-distribution and $n - 2$ degrees of freedom.

Example E7.3

Construct a 95% confidence interval for B for the data in Example E7.1.

SOLUTION:

Refer to the solutions of Examples E7.1 and E7.2.

$$n = 8,\ b = 19.5083,\ SS_{xx} = 4200 \text{ and } s_e = 189.5304$$

The confidence level is 95%

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{189.5304}{\sqrt{4200}} = 2.9245$$

$$df = n - 2 = 8 - 2 = 6$$

$$\alpha/2 = 0.5 - (0.95/2) = 0.025$$

From the *t*-distribution table (Appendix-*G*), the value of *t* for 6 *df* and 0.025 area in the right tail of the *t*-distribution is 2.447.

The 95% confidence interval for *B* is

$$b \pm t\, s_b = 19.5083 \pm (2.447)(2.9245) = 19.5083 \pm 7.1563 = 12.3520 \text{ to } 26.6646$$

Hence, we are 95% confident that the slope *B* of the population regression line is between 12.3520 and 26.6646.

7.10 HYPOTHESIS TESTING ABOUT *B*

To test the hypothesis that *x* does not determine *y* linearly, we test the null hypothesis that the slope of the regression line is zero, that is, $B = 0$. The alternative hypothesis is one of the following:

(*a*) *x* determines *y*, that is, $B \# 0$.

(*b*) *x* determines *y* positively, that is, $B > 0$.

(*c*) *x* determines *y* negatively, that is, $B < 0$.

The value of the *test statistic t* for *b* is calculated as

$$t = \frac{b - B}{s_b} \tag{7.31}$$

The value of *B* is substituted from the null hypothesis.

Example E7.4

Test at the 1% significance level if the slope of the regression line for the Example E7.1 is positive.

SOLUTION:

Refer to the solutions obtained earlier for Examples E7.1 to E7.3.

Here $n = 8,\ b = 19.5083 \text{ and } s_b = 2.9245$

The two hypotheses are

$$H_0: B = 0 \text{ (slope is zero)}$$

$$H_1: B > 0 \text{ (slope is positive)}$$

We can also write the null hypothesis as H_0: $B \leq 0$ (slope is either zero or negative).

Since the sample size is small ($n < 30$) and σ_e is unknown, we use the *t*-distribution to make the test about *B*.

The significance level is 0.01. The > sign in the alternative hypothesis indicates that the test is right-tailed. Hence

Area in the right tail of the *t*-distribution , $\alpha = 0.01$

$$df = n - 2 = 8 - 2 = 6$$

From the t-distribution table (Appendix-G), the critical value of t for 6 df and 0.01 area in the right tail of the t-distribution is 3.143, as shown in Fig. E7.4.

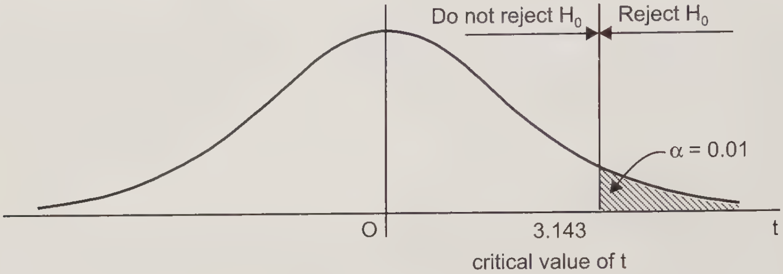


Fig. E7.4

Now $$t = \frac{b - B}{s_b} = \frac{19.5083 - 0 \text{ (from } H_0)}{2.9245} = 6.6706$$

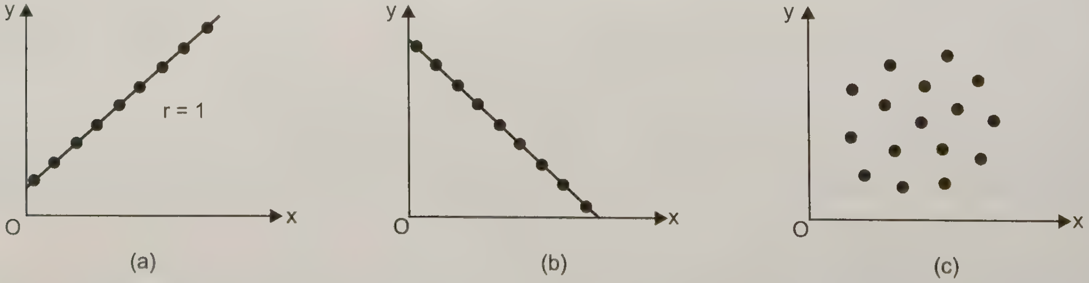
The value of the test statistic $t = 6.6706$ is greater than the critical value of $t = 3.143$, and it falls in the rejection region. Therefore, we reject the null hypothesis and conclude that x determines y positively.

7.11 LINEAR CORRELATION

Linear correlation coefficient is a measure of the relationship between two variables. Linear correlation coefficient measures how closely the points in a scatter diagram are spread around the regression line. The correlation coefficient calculated for the population is denoted by ρ and the one calculated for sample data is denoted by r. The linear correlation coefficient r measures the strength of the linear relationship between the paired x- and y- quantitative values in a sample. The linear correlation coefficient is sometimes referred to as the Pearson product moment correlation coefficient in honor of Karl Pearson (1857–1936), who originally developed it. Square of the correlation coefficient is equal to the coefficient of determination. The value of the *correlation coefficient* always lies in the range –1 to 1. Hence $-1 \le \rho \le 1$ and $-1 \le r \le 1$.

If $r = 1$, it refers to a case of *perfect positive linear correlation* and all points in the scatter diagram lie on a straight line that slopes upward from left to right, as shown in Fig. 7.13. If $r = -1$, the correlation is said to be *perfect negative linear correlation* and all points in the scatter diagram fall on a straight line that slopes downward from left to right, as shown in Fig.7. 13(b).

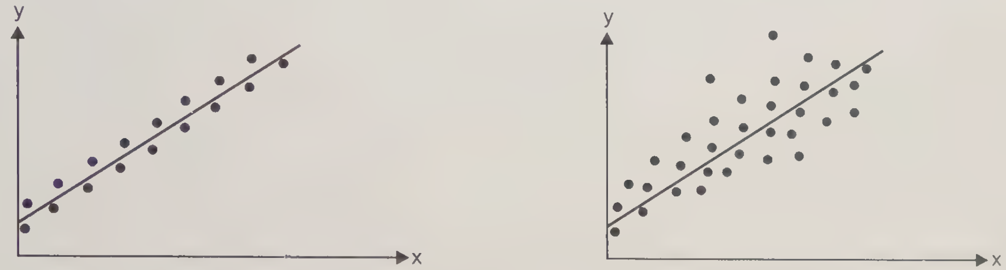
When there is *no linear correlation* between the two variables and r is close to 0. Also, in this case, all the points are scattered all over the diagram as shown in Fig. 7.13(c).



(a) (b) (c)

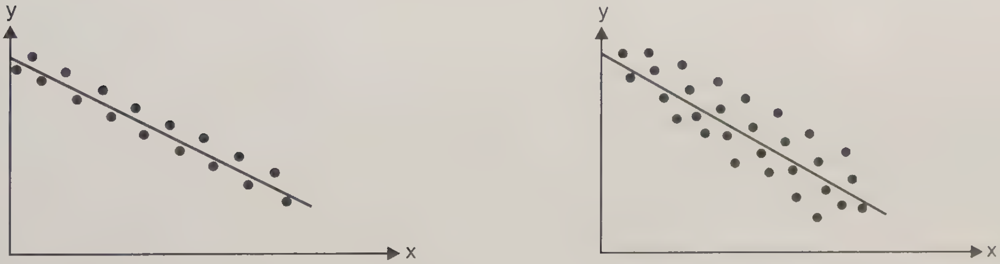
Fig. 7.13: Linear correlation between two variables (a) Perfect positive linear correlation, $r = 1$, (b) Perfect negative linear correlation, $r = -1$, (c) No linear correlation, $r = 0$

Two variables are said to have a *strong positive linear correlation* when the correlation is positive and close to 1. If the correlation between the two variables is positive but close to zero, then the variables have a *weak positive linear correlation*. Similarly, when the correlation between two variables is negative and close to –1, then the variables are said to have a *strong negative linear* correlation. A *weak negative linear correlation* exists when the correlation between the variables is negative but close to zero. The above four cases are shown in Figs. 7.14(*a*) to (*d*).



(a) Strong positive correlation
(r close to 1)

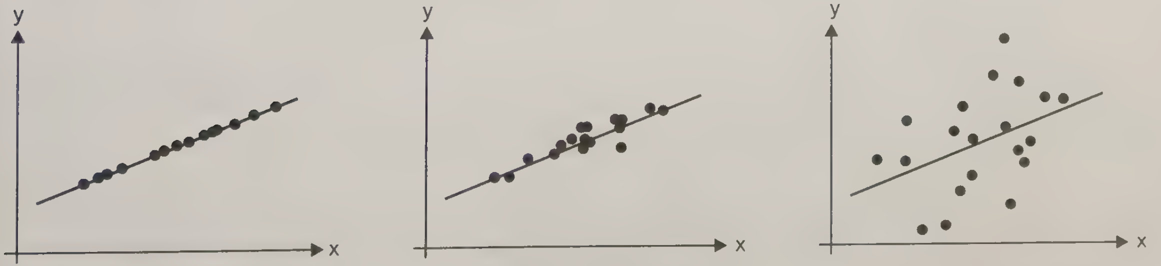
(b) Weak positive linear correlation
(r is positive but close to zero)



(c) Strong negative linear correlation
(r close to –1)

(d) Weak negative linear correlation
(r is negative and close to zero)

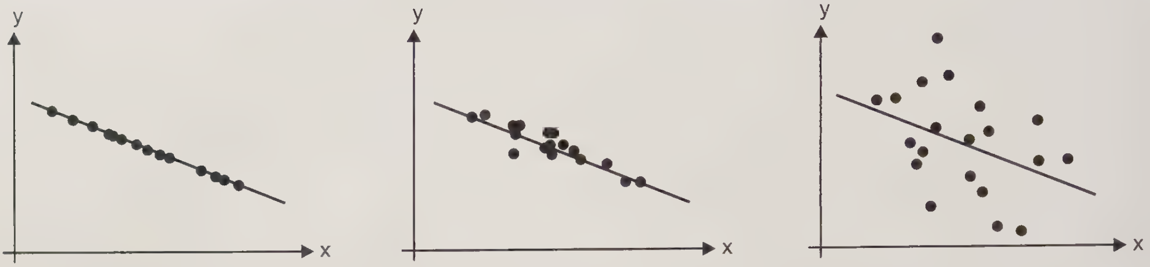
Fig. 7.14: Linear correlation between variables



(a) Perfect positive linear correlation r = 1

(b) Strong positive linear correlation r = 0.92

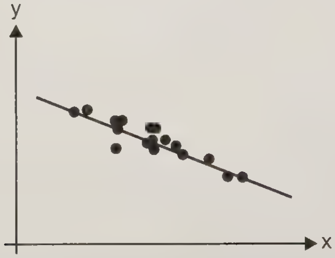
(c) Weak positive linear correlation r = 0.42



(d) Perfect negative linear correlation r = 1

(e) Strong negative linear correlation r = 0.92

(f) Weak negative linear correlation r = 0.42



(g) No linear correlation (linearly uncorrelated) r = 0

Fig. 7.14(a): Various degrees of linear correlation

The *simple linear correlation,* denoted by r, measures the strength of the linear relationship between two variables for a sample and is calculated as

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} \tag{7.32}$$

It should be noted here that r and b calculated for the same sample will always have the same sign.

Properties of the Linear Correlation Coefficient r

1. The values of r is always between –1 and +1 inclusive. That is, $-1 \le r \le 1$.
2. The values of r does not change if all values of either variable are converted to a different scale.
3. The value of r is not affected by the choice of x or y. Interchange all x- and y-values and the value of r will not change.
4. r measures the strength of a linear relationship. The magnitude of r indicates the strength of the linear relationship. A value of r close to –1 or to 1 indicates a strong linear relationship between the variables and that the variable x is a good linear predictor of the variable y. That is, the regression equation is extremely useful for making predictions. A value of r near 0 indicates at most a weak linear relationship between the variables and that the variable x is a poor linear predictor of the variable y. That is, the regression equation is either useless or not very useful for making predictions. It is not designed to measure the strength of a relationship that is not linear.
5. The sign of r suggests the type of linear relationship. A positive value of r suggests that the variables are positively correlated, meaning that y tends to increase linearly as x increases, with

the tendency being greater the closer that r is to 1. A negative value of r suggests that the variables are negatively linearly correlated, meaning that y tends to decrease linearly as x increases, with the tendency being greater the closer that r is to –1.

6. r reflects the slope of the scatter plot. The linear correlation coefficient is positive when the scatter plot shows a positive slope and is negative when the scatter plot shows a negative slope.
7. The sign of r and the sign of the slope of the regression line are identical. If r is positive, so is the slope of the regression line. That is, the regression line slopes upward. If r is negative, so is the slope of the regression line and the regression line slopes downward.

Explained and Unexplained Variation

The total variation is defined as $\Sigma(y-\bar{y})^2$, i.e., the sum of the squares of the deviations of the values of y from the mean $\bar{y}$. This can be written as

$$\Sigma(y-\bar{y})^2 = \Sigma(y-\hat{y})^2 + \Sigma(\hat{y}-\bar{y})^2 \tag{7.32a}$$

where $\hat{y}$ is the value of y for given values of x as estimated from $\hat{y} = a + bx$, a measure of the scatter about the regression line of y on x.

The first term on the right side of Eq. (7.32a) is called the *unexplained variation* while the second term is called the *explained variation*. The deviations $\hat{y}-\bar{y}$ have a definite pattern while the deviations $y-\hat{y}$ behave in a random or unpredictable manner. Similar results hold true for the variable x.

The ratio of the explained variation to the total variation is called the *coefficient of determination*. If there is zero explained variation i.e., the total variation is all unexplained, then this ratio is zero. If there is zero unexplained variation i.e., the total variation is all explained, the ratio is one. In all other cases, the ratio lies between zero and one. The ratio is always non-negative.

The quantity, r, is called the coefficient of correlation, and it is given by

$$r = \pm\sqrt{\frac{\text{explained variation}}{\text{total variation}}} = \pm\sqrt{\frac{\Sigma(\hat{y}-\bar{y})^2}{\Sigma(y-\bar{y})^2}}\Bbbk \tag{7.32b}$$

r varies between –1 and +1. The signs ± are used for positive linear correlation and negative error correlation respectively. r is a dimensionless quantity. The coefficient of determination equals the square of the linear correlation coefficient.

Example E7.5

Determine the correlation coefficient for the data given in Example E7.1.

SOLUTION:

Refer to the solutions obtained earlier for Examples E7.1 and E7.2. We have

$$SS_{xy} = 81935,\ SS_{xx} = 4200 \text{ and } SS_{yy} = 1813945.875$$

Hence,

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}} = \frac{81935}{\sqrt{(4200)(1813945.875)}} = 0.9387$$

7.12 ESTIMATING THE MEAN VALUE OF *y*

The population regression model is

$$y = A + Bx + \in$$

and the mean value of y for a given x is denoted by $\mu_{y/x}$. Since the mean value of $\in$ is assumed zero, the mean value of y is given by

$$\mu_{y/x} = A + Bx$$

The value of $\hat{y}$ obtained from the sample regression line by substituting the value of x, is the *point estimate* of $\mu_{y/x}$ for that x. All possible samples of the same size taken from the same population will give different regression lines as shown in Fig. 7.15 and a different point estimate of $\mu_{y/x}$. Hence, a confidence interval constructed for $\mu_{y/x}$ based on one sample will give a more reliable estimate of $\mu_{y/x}$ than with a point estimate.

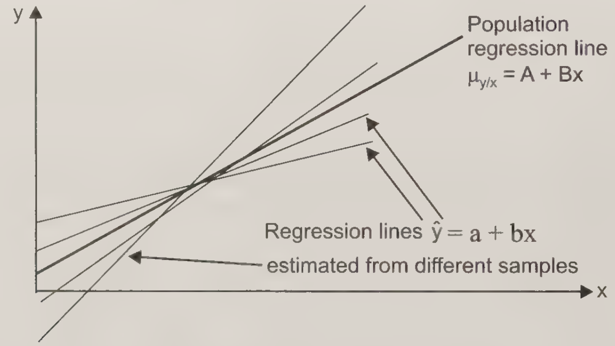

y
Population regression line $\mu_{y/x}$ = A + Bx
Regression lines $\hat{y}$ = a + bx estimated from different samples
x

Fig. 7.15: Population and sample regression lines

In order to construct a confidence interval for $\mu_{y/x}$, the mean, the standard deviation, and the shape of the sampling distribution of its point estimator $\hat{y}$ should be known. The point estimator $\hat{y}$ of $\mu_{y/x}$ is assumed normally distributed with a mean of $A + Bx$ and a standard deviation of

$$\sigma_{\hat{y}_m} = \sigma_{\in} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} \tag{7.33}$$

where $\sigma_{\hat{y}_m}$ = the standard deviation $\hat{y}$ of when it is used to estimate $\mu_{y/x}$

x_0 = the value of x for which we are estimating $\mu_{y/x}$

$\sigma_{\in}$ = the population standard deviation of $\in$.

Here, $\sigma_{\in}$ is not known and it is estimated by the standard deviation of sample errors s_e. Therefore, $\sigma_{\in}$ will be replaced by s_e and $\sigma_{\hat{y}_m}$ by in Eq. (7.33). Also, we use the normal distribution for a sample size which is large ($n \geq 30$) to make a confidence interval for $\mu_{y/x}$. Similarly, t-distribution will be used for a small sample size ($n < 30$).

The $(1 - \alpha)$ 100% confidence interval for $\mu_{y/x}$ for $x = x_0$ is given by

$$\hat{y} \pm t\, s_{\hat{y}_m} \tag{7.34}$$

where the value of t is obtained from the t-distribution table (Appendix-G) for $\alpha/2$ area in the right tail of the t distribution curve and $df = n - 2$. The value of $s_{\hat{y}_m}$ is computed from

$$s_{\hat{y}_m} = s_e\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} \tag{7.35}$$

Example E7.6

Refer to Example E7.1 for the velocity and forces. Determine a 99% confidence interval for the mean force with a velocity of 50 m/s.

Solution:

From the solution of Example E7.1, the point estimate of the mean force for $x = 50$ m/s is

$$\hat{y} = -236.50 + 19.5083x \text{ (N)} = -236.50 + 19.5083(50) = 738.9167$$

The confidence level is 99%. Therefore, the area in each tail of the t-distribution is

$$\alpha/2 = 0.5 - (0.99/2) = 0.005$$

The degrees of freedom are

$$df = n - 2 = 8 - 2 = 6$$

From the t-distribution table (Appendix-G), the t value for 0.005 area in the right tail of the t-distribution and 6 df is 3.707.

From the solutions of Examples E7.1 and E7.2, we have

$$s_e = 189.5304, \ \bar{x} = 45 \text{ and } SS_{xx} = 4200$$

The standard deviation of as an estimate of $\mu_{y/x}$ for $x_0 = 50$ m/s is given by

$$s_{\hat{y}_m} = s_e\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}} = 189.5304\sqrt{\frac{1}{8} + \frac{(50 - 45)^2}{4200}} = 68.5860$$

Therefore, the 99% confidence interval for $\mu_{y/x}$ is

$$\hat{y} \pm ts_{\hat{y}_m} = 738.9167 \pm 3.707(68.5860) = 738.9167 \pm 254.2483 = 484.6684 \text{ to } 993.1650$$

Hence, with 99% confidence we can state that the mean force for all velocities with a velocity of 50 m/s is between 484.6684 and 993.1650 N.

7.13 ESTIMATING A PARTICULAR VALUE OF *y*

The regression model can also be used to estimate a particular value of y for a given value of x, say x_0. Here, a confidence interval constructed for y_p based on one sample can give a more reliable estimate of y_p than with a point estimate. The confidence interval constructed for y_p is more commonly called a *prediction interval.*

The procedure to construct a prediction interval for y_p is similar to that for constructing a confidence interval for $\mu_{y/x}$. However, the standard deviation of $\hat{y}$ is larger since we predict a single value of y than we estimate $\mu_{y/x}$.

The point estimator $\hat{y}$ of y_p is normally distributed with a mean of $A + Bx$ and a standard deviation of

$$\sigma_{\hat{y}_p} = \sigma_{\in}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}} \tag{7.36}$$

where $\sigma_{\hat{y}_p}$ = the standard deviation of the predicted value of y

x_0 = the value of x for which the prediction of y is made

$s_{\in}$ = the population standard deviation of $\in$.

In general, $\sigma_{\in}$ is not known. We use the normal distribution in order to make a prediction interval for y_p when the sample size is large ($n \geq 30$) and t-distribution when the sample size is small ($n < 30$).

The $(1 - \alpha)$ 100% *prediction interval* for the predicted value of y, denoted by y_p for $x = x_0$ is given by

$$\hat{y} \pm t_{\hat{y}_p} \tag{7.37}$$

where the value of t is obtained from the t-distribution table (Appendix-G) for $\alpha/2$ area in the right tail of the t-distribution curve and $df = n - 2$. The value of $s_{\hat{y}_p}$ is computed as

$$s_{\hat{y}_p} = s_e\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}} \tag{7.38}$$

Example E7.7

Refer to Example E7.1 on velocities and forces. Find a 99% prediction interval for the predicted forces for a randomly selected velocity of 50 m/s.

SOLUTION:

The regression line was estimated in Example E7.1. The point estimate of the predicted force for $x = 50$ m/s is given by from Example E7.6

$$\hat{y} = -236.50 + 19.5083x \text{ (N)} = 738.9167$$

The area in each tail of the t-distribution for a 99% prediction level is $\alpha/2 = 0.5 - 0.99/2 = 0.005$.

The degrees of freedom are

$$df = n - 2 = 8 - 2 = 6$$

From the t-distribution table (Appendix-G), the t value for 0.005 area in the right tail of the t-distribution curve and 6 df is 3.707. From the results of Examples E7.1 and E7.2, we have

$$s_e = 189.5304, \quad \bar{x} = 45 \text{ and } SS_{xx} = 4200$$

The standard deviation $\hat{y}$ of as an estimator of y_p for $x = 45$ m/s is computed using

$$s_{\hat{y}_p} = s_e\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}} = (189.5304)\sqrt{1+\frac{1}{8}+\frac{(50-45)^2}{(4200)}} = 201.5584$$

Therefore, the 99% prediction interval for y_p for $x = 50$ m/s is

$$\hat{y} \pm t\, s_{\hat{y}_p} = 738.9167 \pm (3.707)(201.5584) = 738.9167 \pm 747.1770 = -8.2603 \text{ to } 1486.094$$

Hence, with 99% confidence we can state that the predicted force for a velocity of 50 m/s is between –8.2603 and 1486.094. We note here that, this interval is much wider than the one for the mean value of y for $x = 50$ m/s computed from Example E7.6, which was 484.6684 to 993.1650 (N). This always happens to be true and in general the prediction interval for predicting a single value of y is always larger than the confidence interval for estimating the mean value of y for a certain value of x.

7.14 LINEARIZATION OF NON-LINEAR RELATIONSHIPS

Linear regression provides a powerful technique for fitting a best line to data. There exists many situations in science and engineering that show the relationship between the quantities that are being considered is not linear. There are several examples of non-linear functions used for curve fitting. A few of them were described in Table 7.1.

Non-linear regression techniques are available to fit these equations in Table 7.1 to data directly. A simpler alternative is to use analytical manipulations to transform the equations into a linear form. Then linear regression can be used to fit the equations to data.

For instance, $y = bx^m$ can be linearized by taking its natural logarithm to give

$$\ell n\, y = \ell n\, b + m\, \ell n\, x \tag{7.39}$$

A plot of $\ell n\, y$ versus $\ell n\, x$ will give a straight line with a slope of m and an intercept of $\ell n\, b$ as shown in Fig. 7.16.

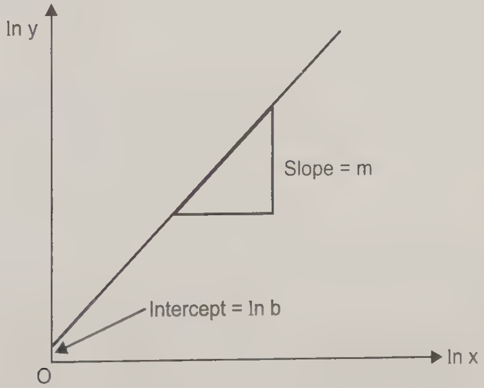


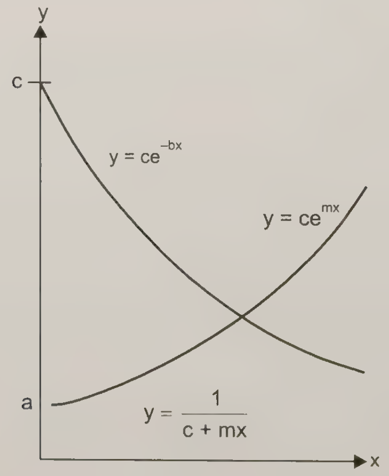
Fig. 7.16: Linearized version of the exponential equation $y = bx^m$

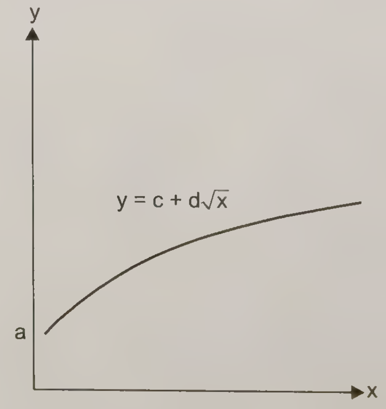
Many other non-linear equations can be transformed into linear form in a similar way. Table 7.2 lists several such equations.

Table 7.2

No.	Non-linear equation	Linear form	Relationship to $\hat{y} = a + b\hat{x}$	Values for least-squares regression
1.	$y = cx^m$	$\ell n(y) = m\ell n(x) + \ell n(c)$	$\hat{y} = \ell n(y),\ \hat{x} = \ell n(x)$ $b = m, a = \ell n(c)$	$\ell n(x_i)$ and $\ell n(y_i)$
2.	$y = ce^{mx}$	$\ell n(y) = mx + \ell n(c)$	$\hat{y} = \ell n(y),\ \hat{x} = x$ $b = m, a = \ell n(c)$	x_i and $\ell n(y_i)$
3.	$y = c\ 10^{mx}$	$\log(y) = mx + \log c$	$\hat{y} = \log(y),\ \hat{x} = x$ $b = m, a = \log(c)$	x_i and $\ell n(y_i)$
4.	$y = \dfrac{1}{mx + c}$	$\dfrac{1}{y} = mx + c$	$\hat{y} = \dfrac{1}{y},\ \hat{x} = x$ $b = m, a = c$	x_i and $\dfrac{1}{y_i}$
5.	$y = \dfrac{mx}{c + x}$	$\dfrac{1}{y} = \dfrac{c}{mx} + \dfrac{1}{m}$	$\hat{y} = \dfrac{1}{y},\ \hat{x} = \dfrac{1}{x}$ $b = \dfrac{c}{m},\ a = \dfrac{1}{m}$	$\dfrac{1}{x_i}$ and $\dfrac{1}{y_i}$
6.	$xy^c = d$ Gas equation	$\log y = \dfrac{1}{c}\log d = \dfrac{1}{c}\log x$	$\hat{y} = \log y,\ \hat{x} = \log x$ $a = \dfrac{1}{c}\log d,\ b = -\dfrac{1}{c}$	$\log x_i$ and $\log y_i$
7.	$y = cd^x$	$\log y = \log c + x \log d$	$\hat{y} - \log y,\ \hat{x} = x$ $a = \log c, b = \log d$	x_i and $\log y_i$
8.	$y = c + d\sqrt{x}$	$y = c + d\hat{x}$ where $\hat{x} = \sqrt{x}$	$\hat{y} = y$ and $\hat{x} = \sqrt{x}$ $a = c$ and $b = d$	$\sqrt{x_i}$ and y_i

The curves in Fig. 7.16(a) may be used as guides to some of the simpler variable transformations.





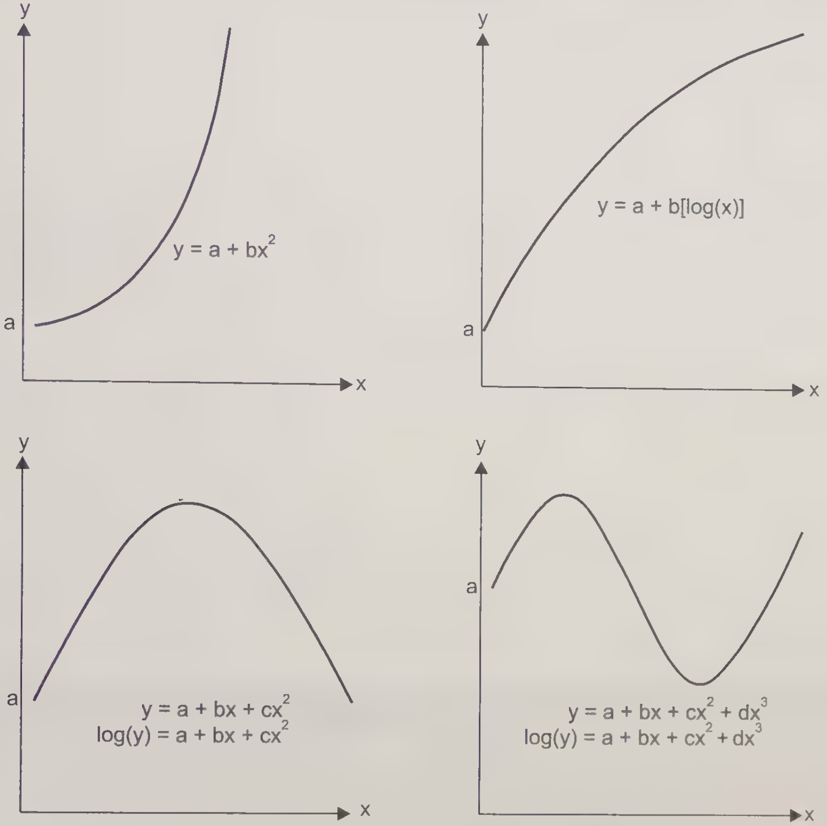


Fig. 7.16(a): Non-linear data curves

Example E7.8

Fit $y = cx^m$ (power function) to the data in Example E7.1 using a logarithmic transformation.

SOLUTION:

The data can be set up in tabular form as shown in Table E7.8.

Table E7.8

i	x_i	y_i	$\log x_i$	$\log y_i$	$(\log x_i)^2$	$(\log x_i)(\log y_i)$
1	10	24	1.0000	1.3802	1.0000	1.3802
2	20	68	1.3010	1.8325	1.6927	2.3841
3	30	378	1.4771	2.5775	2.1819	3.8073
4	40	552	1.6021	2.7419	2.5666	4.3928
5	50	608	1.6990	2.7839	2.8865	4.7298
6	60	1218	1.7782	3.0856	3.1618	5.4867
7	70	831	1.8451	2.9196	3.4044	5.3870
8	80	1452	1.9031	3.1620	3.6218	6.0175
Σ	360	5131	12.606	20.483	20.516	33.585

The means are computed as

$$\bar{x} = \frac{\Sigma \log x}{n} = \frac{12.606}{8} = 1.5757$$

$$\bar{y} = \frac{\Sigma \log y}{n} = \frac{20.483}{8} = 2.5604$$

The slope and the intercept are then calculated using Eqs. (7.13), (7.14), (7.15) and (7.16).

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{n\Sigma(\log x_i)(\log y_i) - (\Sigma \log x_i)(\Sigma \log y_i)}{n\Sigma \log x_i^2 - (\Sigma \log x_i)^2} = \frac{8(33.585) - (12.606)(20.483)}{8(20.516) - (12.606)^2} = 2.0055$$

$$a = \bar{y} - b\bar{x} = 2.5604 - 2.0055(1.5757) = -0.5997$$

The least-square fit is

$$\log y = -0.5997 + 2.0055 \log x$$

The fit along with the data is shown in Fig. E7.8.

Transforming to the original coordinates, we have

$$c = 10^{-(0.5997)} = 0.2514 \text{ and } m = 2.0055$$

Hence the least-squares fit is

$$y = 0.2514\,x^{2.0055}$$

7.15 POLYNOMIAL REGRESSION

The least-squares procedure described in earlier sections can be readily extended to fit the data to a higher order polynomial. Consider a second-order polynomial or quadratic:

$$y = a + bx + cx^2 + e \tag{7.40}$$

The sum of the squares of the residuals is

$$S_r = \sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2)^2 \tag{7.41}$$

To generate the least-squares fit, we take the first derivative of Eq. (7.41) with respect to each of the unknown coefficients of the polynomial.

$$\begin{aligned} \frac{\partial S_r}{\partial a} &= -2\sum_{i=1}^{n} (y_i - a - bx_i - cx_i^2) \\ \frac{\partial S_r}{\partial b} &= -2\sum_{i=1}^{n} x_i(y_i - a - bx_i - cx_i^2) \\ \frac{\partial S_r}{\partial c} &= -2\sum_{i=1}^{n} x_i^2(y_i - a - bx_i - cx_i^2) \end{aligned} \tag{7.42}$$

Equations (7.42) are set equal to zero and rearranged to obtain the following set of normal equations:

$$na + \left(\sum_{i=1}^{n} x_i\right) b + \left(\sum_{i=1}^{n} x_i^2\right) c = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i\right)a+\left(\sum_{i=1}^{n} x_i^2\right)b+\left(\sum_{i=1}^{n} x_i^3\right)c=\sum_{i=1}^{n} x_i y_i \tag{7.43}$$

$$\left(\sum_{i=1}^{n} x_i^2\right)a+\left(\sum_{i=1}^{n} x_i^3\right)b+\left(\sum_{i=1}^{n} x_i^4\right)c=\sum_{i=1}^{n} x_i^2 y_i$$

Equation (7.43) are all linear equations in three unknowns: a, b and c. These coefficients a, b and c can be determined directly from the observed data.

The above procedure can be easily extended to an m^{th} order polynomial as in

$$y=a+bx+cx^2+dx^3+\cdots+zx^m+e \tag{7.44}$$

Hence, the determination of the coefficients of an m^{th} order polynomial is equivalent to solving a system of ($\mu + 1$) simultaneous linear equations.

The standard error in this case is given by

$$S_{y/x}=\sqrt{\frac{S_r}{n-(m+1)}} \tag{7.45}$$

The coefficient of determination, r^2, can be computed for a polynomial regression with Equation (7.25).

Example E7.9

Fit a second-order polynomial to the data in Table E7.9 and determine the total standard deviation, the standard error of the estimate and the correlation coefficient.

Table E7.9

x_i	0	1	2	3	4	5
y_i	2	8	14	27	41	61

SOLUTION:

Tables E7.9(a) and (b) shows the computations for an error analysis of the quadratic least-squares fit.

Table E7.9(a)

	x_i	y_i	$x_i y_i$	x_i^2	$x_i^2 y_i$	x_i^3	x_i^4
	0	2	0	0	0	0	0
	1	8	8	1	8	1	1
	2	14	28	4	56	8	16
	3	27	81	9	243	27	81
	4	41	164	16	656	64	256
	5	61	305	25	1525	125	625
Σ	15	153	586	55	2488	225	979

Table E7.9(b)

	x_i	y_i	$(y_i-\bar{y})^2$	$(y_i-a-bx_i-cx_i^2)^2$
	0	2	552.3	0.2500
	1	8	306.3	1.3391
	2	14	132.3	0.6862
	3	27	2.3	0.2951
	4	41	240.3	0.5300
	5	61	1260.3	0.1282
Σ	15	153	2493.50	3.2286

Hence, the simultaneous linear equations are

$$\begin{bmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{bmatrix}\begin{Bmatrix} a \\ b \\ c \end{Bmatrix}=\begin{Bmatrix} 153 \\ 586 \\ 2488 \end{Bmatrix}$$

Refer to Appendix-*I* (Cramer's rule for solving a system of linear algebraic equations).

Here
$$D = \begin{vmatrix} 6 & 15 & 55 \\ 15 & 55 & 225 \\ 55 & 225 & 979 \end{vmatrix} = 3920;$$

$$D_1 = \begin{vmatrix} 153 & 15 & 55 \\ 586 & 55 & 225 \\ 2488 & 225 & 979 \end{vmatrix} = 9800;$$

$$D_2 = \begin{vmatrix} 6 & 153 & 55 \\ 15 & 586 & 225 \\ 55 & 2488 & 979 \end{vmatrix} = 9884;$$

$$D_3 = \begin{vmatrix} 6 & 15 & 153 \\ 15 & 55 & 586 \\ 55 & 225 & 2488 \end{vmatrix} = 7140$$

Therefore,

$$a = \frac{D_1}{D} = \frac{9800}{3920} = 2.5$$

$$b = \frac{D_2}{D} = \frac{9884}{3920} = 2.5214$$

$$c = \frac{D_3}{D} = \frac{7140}{3920} = 1.8214$$

These equations can also be solved to determine the coefficients using MATLAB:

```
>>A = [6 15 55; 15 55 225; 55 225 979];
>>b = [153; 586; 2488];
x = A/b
x = 2.5  2.5214  1.8214
```

or $a = 2.5, b = 2.5214$, and $c = 1.8214$

Hence, the least-squares quadratic equation for this problem is

$$y = 25 + 2.5214x + 1.8214x^2$$

The standard error of the estimate is based on the regression polynomial given by Eq. (7.48), where $S_r = \Sigma(y - \hat{y})^2$. Here, we have

$$S_{y/x} = \sqrt{\frac{S_r}{n-(m+1)}} = \sqrt{\frac{3.229}{6-(2+1)}} = 1.0374$$

The coefficient of determination is given by Eq. (7.25)

$$r^2 = \frac{S_t - S_r}{S_t} = \frac{2493.5 - 3.229}{2493.5} = 0.9987$$

where $$S_t = \Sigma(y_i - \bar{y})^2$$

and $$S_r = \Sigma(y_i - \hat{y})^2$$

Therefore, the correlation coefficient $r = \sqrt{0.9987} = 0.99935$. These results show that 99.935% of the original uncertainty has been explained by the model.

7.16 QUANTIFICATION OF ERROR OF LINEAR REGRESSION

Noting that the sum of the squares is defined as

$$S_r = \sum_{i=1}^{n} (y_i - a - bx_{1,i} - cx_{2,i})^2 \tag{7.46}$$

Equation (7.46) is similar to the equation

$$S_t = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{7.47}$$

In Eq. (7.47), the squares of the residual represented the squares of the discrepancy between the data and a single estimate of the measure of central tendency (the mean). The squares of the residual represent the squares of the vertical distance between the data and another measure of central tendency (the straight line). If the spread of the points around the line is of similar magnitude along the entire range of data and the distribution of these points about the line is normal, then the least-squares regression will provide the best estimates of a and b. This is known as the *maximum likelihood principle*. Also, if these criteria are met, a standard deviation for the regression line can be determined as

$$S_{y/x} = \sqrt{\frac{S_r}{n-2}} \tag{7.48}$$

where $S_{y/x}$ is called the *standard error of the estimate.*

The difference between the S_t and S_r quantifies the improvement or error reduction due to describing the data in terms of a straight line rather than as an average value. The difference is therefore normalized to S_t to give

$$r^2 = \frac{S_t - S_r}{S_t} \tag{7.49}$$

where r^2 is called *the coefficient of determination* and r is the *correlation coefficient*. For a perfect fit, $S_r = 0$, and $r^2 = 1$, indicating that the line explains 100% of the variability of the data. For $r^2 = 0$, $S_r = S_t$ and the fit represents no improvement. An alternative formulation for r is given by

$$r = \frac{n\sum_{i=1}^{n}(x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

$$= \frac{\sum_{i=1}^{n}(x_i y_i) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{\left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}} \tag{7.50}$$

Example E7.10

Determine (*a*) the total standard deviation, (*b*) the standard error of the estimate and (*c*) the correlation coefficient for the data in Example E7.1.

SOLUTION:

Table E7.10 shows the data and summation to compute the goodness-of-fit statistics. From Example E7.1, we have $a = -236.50$, $b = 19.5083$, $\bar{x} = 45$ and $\bar{y} = 641.375$.

Table E7.10

i	x_i	y_i	$a + bx_i$	$(y_i - \bar{y})^2$	$(y_i - a - bx_i)^2$
1	10	24	–41.4167	381151.8906	4279.3403
2	20	68	153.6667	328758.8906	7338.7778
3	30	378	348.7500	69366.3906	855.5625
4	40	552	543.8333	7987.8906	66.6944
5	50	608	738.9167	1113.8906	17139.1736
6	60	1218	934.0000	332496.3906	80656.0000
7	70	831	1129.0833	35957.6406	88853.6736
8	80	1452	1324.6667	657112.8906	16341.3611
Σ	360	5131	5131	1813945.875	215530.583

The standard deviation is given by

$$S_y = \sqrt{\frac{S_t}{n-1}}$$

where S_t is the total sum of the squares of the residuals between the data points and the mean.

Hence $$S_y = \sqrt{\frac{1813945.875}{8-1}} = 476.1746$$

The standard error of the estimate is

$$S_{y/x} = \sqrt{\frac{S_r}{n-2}} = \sqrt{\frac{215530.583}{8-2}} = 164.1320$$

Since $S_{y/x} < S_y$, the linear regression model has merit.

The coefficient of determination r^2 is given by Equation (7.49)

$$r^2 = \frac{S_t - S_r}{S_t} = \frac{1813945.875 - 215530.583}{1813945.875} = 0.8812$$

or
$$r = \sqrt{0.8812} = 0.9387$$

These results indicate that 93.87% of the original uncertainty has been explained by the linear model.

7.17 MULTIPLE LINEAR REGRESSION

Consider a function y which is a linear function of x_1 and x_2 as in

$$y = a + bx_1 + cx_2 + e \tag{7.51}$$

Equation (7.51) is quite useful in fitting experimental data where variable being studied is often a function of two other variables. For this two-dimensional case, the regression *line* becomes a *plane*. The best values of the coefficients are obtained by formulating the sum of the squares of the residuals:

$$S_r = \sum_{i=1}^{n} (y_i - a - bx_{1,i} - cx_{2,i})^2 \tag{7.52}$$

Differentiating equation (7.52) with respect to each of the unknown coefficients, we get

$$\frac{\partial S_r}{\partial a} = -2\sum_{i=1}^{n} (y_i - a - bx_{1,i} - cx_{2,i})$$

$$\frac{\partial S_r}{\partial b} = -2\sum_{i=1}^{n} x_{1,i}(y_i - a - bx_{1,i} - cx_{2,i})$$

$$\frac{\partial S_r}{\partial c} = -2\sum_{i=1}^{n} x_{2,i}(y_i - a - bx_{1,i} - cx_{2,i})$$

The coefficient giving the minimum sum of the squares of the residuals are obtained by setting the partial derivatives equal to zero and expressing the result in matrix form as

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{1,i} & \sum_{i=1}^{n} x_{2,i} \\ \sum_{i=1}^{n} x_{1,i} & \sum_{i=1}^{n} x_{1,i}^2 & \sum_{i=1}^{n} x_{1,i}x_{2,i} \\ \sum_{i=1}^{n} x_{2,i} & \sum_{i=1}^{n} x_{1,i}x_{2,i} & \sum_{i=1}^{n} x_{2,i}^2 \end{bmatrix} \begin{Bmatrix} a \\ b \\ c \end{Bmatrix} = \begin{Bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{1,i}y_i \\ \sum_{i=1}^{n} x_{2,i}y_i \end{Bmatrix} \tag{7.53}$$

Example E7.11

The following data was generated from the equation $y = 7 + 3x_1 + 4x_2$. Use multiple linear regressions to fit this data.

x_1	0	1	2	1	4	7	2	0
x_2	0	2	1	3	5	1	3	4
y	7	18	17	22	39	32	25	23

SOLUTION:

Table E7.11(a)

	y_i	$x_{1,i}$	$x_{2,i}$	$x_{1,i}^2$	$x_{1,i}\,x_{2,i}$	$x_{2,i}^2$	$x_{2,i}\,y_i$	$x_{1,i}\,y_i$
	7	0	0	0	0	0	0	0
	18	1	2	1	2	4	36	18
	17	2	1	4	2	1	17	34
	22	1	3	1	3	9	66	22
	39	4	5	16	20	25	195	156
	32	7	1	49	7	1	32	224
	25	2	3	4	6	9	75	50
	23	0	4	0	0	16	92	0
Σ	183	17	19	75	40	65	513	504

The summations required for Eq. (7.53) are computed in Table E7.11(a) as shown above. Substituting these values in Equation (7.53), we get

$$\begin{bmatrix} 8 & 17 & 19 \\ 17 & 75 & 40 \\ 19 & 40 & 65 \end{bmatrix} \begin{Bmatrix} a \\ b \\ c \end{Bmatrix} = \begin{Bmatrix} 183 \\ 504 \\ 513 \end{Bmatrix}$$

which can be solved using Cramer's rule or MATLAB.

Refer to Appendix-*I* (Cramer's rule for solving a system of linear algebraic equations).

$$D = \begin{vmatrix} 8 & 17 & 19 \\ 17 & 75 & 40 \\ 19 & 40 & 65 \end{vmatrix} = 6180;$$

$$D_1 = \begin{vmatrix} 183 & 17 & 19 \\ 504 & 75 & 40 \\ 513 & 40 & 65 \end{vmatrix} = 43260;$$

$$D_2 = \begin{vmatrix} 8 & 183 & 19 \\ 17 & 504 & 40 \\ 19 & 513 & 65 \end{vmatrix} = 18540;$$

$$D_3 = \begin{vmatrix} 8 & 17 & 183 \\ 17 & 75 & 504 \\ 19 & 40 & 513 \end{vmatrix} = 24720$$

$$a = \frac{D_1}{D} = \frac{43260}{6180} = 7$$

$$b = \frac{D_2}{D} = \frac{18540}{6180} = 3$$

$$c = \frac{D_3}{D} = \frac{24720}{6180} = 4$$

MATLAB Program:

```
>>A = [8 17 19; 17 75 40; 19 40 65];
>>B = [183; 504; 513];
x = A/B
```

$$x = 7; \quad b = 3 \quad \text{and} \quad c = 4$$

or

$$a = 7, \quad b = 3, \quad \text{and} \quad c = 4.$$

which is consistent with the original equation from which the data was derived.

7.18 WEIGHTED LEAST-SQUARES METHOD

Referring to the sections 7.3 and 7.4 and assigning weights w_i to each error, e_i $(i = 1, 2, \ldots, n)$ in Equation (7.4) such that $\Sigma w_i = 1$.

Equation (7.5) can be written as

$$S_r = \sum_{i=1}^{n} w_i [y_i - (a + bx_i)]^2 \tag{7.54}$$

For S_r to be a minimum, we have

$$\frac{\partial S_r}{\partial a} = 0 \text{ and } \frac{\partial S_r}{\partial b} = 0 \tag{7.55}$$

We obtain the normal equations as

$$a + b\Sigma w_i x_i = \Sigma w_i y_i \tag{7.56}$$

$$a\Sigma w_i x_i + b\Sigma w_i x_i^2 = \Sigma w_i x_i y_i \tag{7.57}$$

The solution of Equations (7.56) and (7.57) gives the values for a and b. These values give the minimum of S_r with respect to the weight w_i.

Similarly, for a parabolic equation, we need to minimize

$$S_r = \Sigma w_i [y_i - (a + bx_i + cx_i^2)]^2 \tag{7.58}$$

where $\Sigma w_i = 1$.

For S_r to be minimum, we have

$$\frac{\partial S_r}{\partial a} = 0, \ \frac{\partial S_r}{\partial b} = 0 \text{ and } \frac{\partial S_r}{\partial c} = 0 \tag{7.59}$$

On simplification, we obtain the following normal equations to determine a, b and c.

$$a + b\Sigma w_i x_i + c\Sigma w_i x_i^2 = \Sigma w_i y_i \tag{7.60}$$

$$a\Sigma w_i x_i + b\Sigma w_i x_i^2 + c\Sigma w_i x_i^3 = \Sigma w_i x_i y_i \tag{7.61}$$

$$a\Sigma w_i x_i^2 + b\Sigma w_i x_i^3 + c\Sigma w_i x_i^4 = \Sigma w_i x_i y_i \tag{7.62}$$

7.19 ORTHOGONAL POLYNOMIALS AND LEAST-SQUARES APPROXIMATION

The previous sections considered the problem of least squares approximation to fit a collections of data. This method is also applicable for continuous data.

7.20 LEAST-SQUARES METHOD FOR CONTINUOUS DATA

Let $y = f(x)$ be a continuous function on $[a, b]$ and it is to be approximated by the n^{th} degree polynomial.

$$y = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n \tag{7.63}$$

Here the sum of the squares of residuals S is given by

$$S = \int_a^b w(x)[y - (a_0x + a_2x^2 + \cdots + a_nx^n)]^2\, dx \tag{7.64}$$

where $w(x)$ is a suitable weight function.

The necessary conditions for minimum S are given by

$$\frac{\partial S}{\partial a_0} = \frac{\partial S}{\partial a_1} = \cdots = \frac{\partial S}{\partial a_n} = 0 \tag{7.65}$$

Equation (7.65) give the normal equations as

$$\begin{aligned}
-2\int_a^b w(x)[y - (a_0 + a_1x + a_2x^2 + \cdots + a_nx^n)]\,dx &= 0\\
-2\int_a^b w(x)[y - (a_0 + a_1x + a_2x^2 + \cdots + a_nx^n)]\,x\,dx &= 0\\
-2\int_a^b w(x)[y - (a_0 + a_1x + a_2x^2 + \cdots + a_nx^n)]\,x^2\,dx &= 0\\
\vdots \qquad\qquad &\ \ \vdots\\
-2\int_a^b w(x)[y - (a_0 + a_1x + a_2x^2 + \cdots + a_nx^n)]\,x^n\,dx &= 0
\end{aligned} \tag{7.66}$$

After simplification these equations reduce to

$$\begin{aligned}
a_0\int_a^b w(x)\,dx + a_1\int_a^b xw(x)\,dx + \cdots + a_n\int_a^b x^n w(x)\,dx &= \int_a^b w(x)\,y\,dx\\
a_0\int_a^b xw(x)\,dx + a_1\int_a^b x^2 w(x)\,dx + \cdots + a_n\int_a^b x^{n+1} w(x)\,dx &= \int_a^b w(x)xy\,dx\\
a_0\int_a^b x^2 w(x)\,dx + a_1\int_a^b x^3 w(x)\,dx + \cdots + a_n\int_a^b x^{n+2} w(x)\,dx &= \int_a^b w(x)x^2 y\,dx\\
\vdots \qquad\qquad &\ \ \vdots\\
a_0\int_a^b x^n w(x)\,dx + a_1\int_a^b x^{n+1} w(x)\,dx + \cdots + a_n\int_a^b x^{2n} w(x)\,dx &= \int_a^b w(x)x^n y\,dx
\end{aligned} \tag{7.67}$$

Since $w(x)$ and $y = f(x)$ are known, equation (7.67) form a system of linear equations with $(n + 1)$ unknowns $a_0, a_1, \ldots, a_n$. This system of equations possesses a unique solution. If

$$a_0 = a_0^*, a_1 = a_1^*, \ldots, a_n = a_n^*$$

is the solution for $a_0, a_1, \ldots, a_n$ then the approximate polynomial is given by

$$y = a_0^* + a_1^* x + a_2^* x^2 + \cdots + a_n^* x^n$$

Example E7.12

Construct a least-squares quadrate approximation to the function $f(x) = \sin px$ on [0, 1].

SOLUTION:

The normal equations for $P_2(x) = a_2x^2 + a_1x + a_0$ are

$$a_0\int_0^1 1dx + a_1\int_0^1 x\,dx + a_2\int_0^1 x^2dx = \int_0^1 \sin\pi x\,dx \tag{E.1}$$

$$a_0\int_0^1 x\,dx + a_1\int_0^1 x^2dx + a_2\int_0^1 x^3dx = \int_0^1 x\sin\pi x\,dx \tag{E.2}$$

$$a_0\int_0^1 x^2\,dx + a_1\int_0^1 x^3dx + a_2\int_0^1 x^4dx = \int_0^1 x^2\sin\pi x\,dx \tag{E.3}$$

Performing the integration gives

$$a_0 + \frac{1}{2}a_1 + \frac{1}{3}a_2 = \frac{2}{\pi} \tag{E.4}$$

$$\frac{1}{2}a_0 + \frac{1}{3}a_1 + \frac{1}{4}a_2 = \frac{1}{\pi} \tag{E.5}$$

$$\frac{1}{3}a_0 + \frac{1}{4}a_1 + \frac{1}{5}a_2 = \frac{\pi^2 - 4}{\pi^3} \tag{E.6}$$

Equations (E.4), (E.5) and (E.6)in three unknowns can be solved to obtain

$$a_0 = \frac{12\pi^2 - 120}{\pi^3} \approx -0.050465$$

and

$$a_1 = -a_2 = \frac{720 - 60\pi^2}{\pi^3} \approx 4.12251$$

Consequently, the least squares polynomial approximation of degree 2 for $f(x) = \sin px$ on [0, 1] is $P_2(x) = -4.12251x^2 + 4.12251x - 0.050465$.

7.21 APPROXIMATION USING ORTHOGONAL POLYNOMIALS

In section 7.19, a function is approximated as a polynomial containing the terms $1, x, x^2, \ldots, x^n$. These terms are called *base functions*, since, any function or even discrete data are approximated based on these functions.

Here, we assume that the base functions are some orthogonal polynomials $f_0(x), f_1(x), \ldots, f_n(x)$. Let the given function be approximated as

$$y = a_0f_0(x) + a_1f_1(x) + \cdots + a_nf_n(x) \tag{7.68}$$

where $f_i(x)$ is a polynomial in x of degree i. Then the residue is given by

$$S = \int_a^b w(x)[y - \{a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x)\}]^2 \, dx \tag{7.69}$$

For minimum S, the conditions are given by

$$\frac{\partial S}{\partial a_0} = 0, \frac{\partial S}{\partial a_1} = 0, \cdots, \frac{\partial S}{\partial a_n} = 0 \tag{7.70}$$

Equation (7.70) yield the following normal equations:

$$-2\int_a^b w(x)[y - \{a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x)\}] f_0(x) \, dx = 0$$

$$-2\int_a^b w(x)[y - \{a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x)\}] f_1(x) \, dx = 0$$

$$\vdots \qquad\qquad\qquad\qquad \vdots$$

$$-2\int_a^b w(x)[y - \{a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x)\}] f_n(x) \, dx = 0 \tag{7.71}$$

After simplification, the i^{th} equation can be written as

$$a_0 \int_a^b w(x) f_0(x) f_i(x) \, dx + a_1 \int_a^b w(x) f_1(x) f_i(x) \, dx + \cdots$$

$$+ a_i \int_a^b w(x) f_i^2(x) \, dx + \cdots + a_n \int_a^b w(x) f_n(x) f_i(x) \, dx = \int_a^b w(x) y \, f_i(x) \, dx \tag{7.72}$$

$$i = 0, 1, 2, \ldots, n.$$

A set of polynomial $\{f_0(x), f_1(x), \ldots, f_n(x)\}$ is said to be *orthogonal* with respect to the weight function $w(x)$ if

$$\int_a^b f_i(x) f_j(x) w(x) \, dx = \begin{cases} 0, & \text{if } i \neq j \\ \int_b^a f_i^2(x) w(x) dx, & \text{if } i = j \end{cases} \tag{7.73}$$

Using Equation (7.73), Equation (7.72) can be written as

$$a_i \int_a^b w(x) f_i^2(x) \, dx = \int_a^b w(x) f_i(x) \, dx, \qquad i = 0, 1, 2, \ldots, n$$

Hence,

$$a_i = \frac{\int_a^b w(x) y \, f_i(x) \, dx}{\int_a^b w(x) f_i^2(x) \, dx}, \qquad i = 0, 1, 2, \ldots, n \tag{7.74}$$

From Equation (7.74), we can find the values of $a_0, a_1, \ldots, a_n$ and the least squares approximation is obtained by substituting these values in Equation (7.68). However, the functions $f_0(x), f_1(x), \ldots, f_n(x)$ are unknown. Several orthogonal functions are available in literature. A few of them are given in Table 7.7.

Any one of the orthogonal functions can be selected to fit a function dependent on the given problem.

Table 7.7: Some standard orthogonal polynomials

Name	$f_i(x)$	Interval	$w(x)$
Legendre	$P_n(x)$	$[-1, 1]$	1
Leguerre	$L_n(x)$	$[0, \infty]$	e^{-x}
Hermite	$H_n(x)$	$(-\infty, \infty)$	e^{-x^2}
Chebyshev	$T_n(x)$	$[-1, 1]$	$(1-x^2)^{-1/2}$

7.22 GRAM-SCHMIDT ORTHOGONALIZATION PROCESS

Let $f_i(x)$ be a polynomial in x of degree i and $\{f_i(x)\}$ be a given sequence of polynomials. Then the sequence of orthogonal polynomials over $\{f_i^*(x)\}$ the interval $[a, b]$ with respect to the weight function $w(x)$ can be generated by the following equation

$$f_i^*(x) = x^i - \sum_{r=0}^{i-1} a_{ir} f_r^*(x), \quad i = 1, 2, \ldots, n \tag{7.75}$$

where the constants are a_{ir}, and $f_0^*(x) = 1$.

To obtain a_{ir}, we multiply Equation (7.75) with $w(x)f_n^*(x)$, $0 \le k \le i-1$ and integrating over $[a, b]$, we obtain

$$\int_a^b f_i^*(x) f_n^*(x) w(x)\,dx = \int_a^b x^i f_n^*(x) w(x)\,dx - \int_a^b \sum_{r=0}^{i-1} a_{ir} f_r^*(x) f_n^*(x) w(x)\,dx \tag{7.76}$$

Using the property of orthogonal polynomial, Equation (7.76) becomes

$$\int_a^b x^i f_n^*(x) w(x)\,dx - \int_a^b a_{in} f_n^{*2}(x) w(x)\,dx = 0$$

or

$$a_{in} = \frac{\int_a^b x^i f_n^*(x)\, w(x)\, dx}{\int_a^b f_n^{*2}(x)\, w(x)\, dx}, \qquad 0 \le n \le i-1 \tag{7.77}$$

Hence, the set of orthogonal polynomials $\{f_i^*(x)\}$ are given by

$$f_0^*(x) = 1$$

$$f_i^*(x) = x^i - \sum_{r=0}^{i-1} a_{ir} f_r^*(x), \qquad i = 1, 2, \ldots, n$$

where

$$a_{ir} = \frac{\int_a^b x^i f_n^*(x) w(x)\,dx}{\int_a^b f_n^{*2}(x) w(x)\,dx} \tag{7.78}$$

For the discrete data, the integral is replaced by summation.

Example E7.13

Use Gram-Schmidt orthogonalization process to find the first two orthogonal polynomials on [–1, 1] with respect to the weight function $w(x) = 1$.

SOLUTION:

Let $$f_0^*(x) = 1$$

Hence $$f_1^*(x) = x - a_{10} f_0^*(x)$$

where $$a_{10} = \frac{\int_{-1}^{1} x\,dx}{\int_{-1}^{1} dx} = 0$$

or $$f_1^*(x) = x.$$

The second orthogonal polynomial is given by

$$f_2^*(x) = x^2 - a_{20} f_0^*(x) - a_{21} f_1^*(x)$$

where $$a_{20} = \frac{\int_{-1}^{1} x^2 dx}{\int_{-1}^{1} dx} = \frac{1}{3},$$

$$a_{21} = \frac{\int_{-1}^{1} x^2 \cdot x\,dx}{\int_{-1}^{1} x^2 dx} = 0$$

Hence, $$f_2^*(x) = x^2 - \frac{1}{3} = \frac{1}{3}(3x^2 - 1)$$

Thus, the first two orthogonal polynomials are

$$f_0^*(x) = 1,\ f_1^*(x) = x \text{ and } f_2^*(x) = \frac{1}{3}(3x^2 - 1)$$

7.23 ADDITIONAL EXAMPLE PROBLEMS AND SOLUTIONS

Example E7.14

Determine the equation to the best fitting exponential curve of the form $y = ae^{bx}$ for the data given in Table E7.14.

Table E7.14

x	1	3	5	7	9
y	115	105	95	85	80

SOLUTION:

Refer to Table E7.14(a).

Table E7.14(a)

i	x_i	y_i	$\log y_i$	x_i^2	$x_i \log y_i$
1	1	115	2.0607	1	2.0607
2	3	105	2.0212	9	6.0636
3	5	95	1.9777	25	9.8886
4	7	85	1.9294	49	13.5059
5	9	80	1.9031	81	17.1278
Σ	25	480	9.8921	165	48.6466

Given $y = ae^{bx}$

The normal equations are

$$\Sigma \log y_i = 5A + B\,\Sigma x_i \tag{E.1}$$

$$\Sigma x_i \log y_i = A\,\Sigma x_i + B\,\Sigma\, x_i^2 \tag{E.2}$$

where $A = \log a$ and $B = b \log e$.

Solving the two normal Equations (E.1) and (E.2), we get

$$A = 2.0802 \text{ and } B = -0.0203$$

Hence $a = \text{antilog of } A = 10^{2.0802} = 120.2818$

and $$b = B/\log e = \frac{-0.0203}{\log e} = -0.0075.$$

Hence, the fitted equation is $y = 120.2818\, e^{-0.0075x}$

Example E7.15

For the data given in Table E7.14, find the equation to best fitting curve of the form $y = ab^x$.

Table E7.15

x	1	2	3	4	5
y	130	150	175	190	240

SOLUTION:

The calculations are shown in Table E7.15(a).

Table E7.15(a)

	x	y	log y	x^2	x log y	ŷ (estimated)
	1	130	2.1139	1	2.1139	129.2062
	2	150	2.1761	4	4.3522	149.5433
	3	175	2.2430	9	6.7291	173.0814
	4	190	2.2788	16	9.1150	200.3144
	5	240	2.3802	25	11.9011	231.8555
Σ	15	885	11.192	55	34.2113	884.0108

The normal equations are obtained as follows:

$$y = ab^x \tag{E.1}$$

Taking logarithms (base 10) on both sides of the above Equation (E.1), we get

$$\log y = \log a + x \log b \tag{E.2}$$

or

$$Y = A + Bx \tag{E.3}$$

where $Y = \log y$, $A = \log a$ and $B = \log b$.

Hence, the normal equations are

$$\Sigma Y = nA + B\Sigma x \tag{E.4}$$

$$\Sigma xY = A\Sigma x + B\Sigma x^2 \tag{E.5}$$

Substituting the values from Table E7.13(a) into Equations (E.4) and (E.5), we have

$$11.1920 = 5A + 15B \tag{E.6}$$

$$34.2113 = 15A + 55B \tag{E.7}$$

Solving Equations (E.6) and (E.7), we obtain

$$A = 2.0478 \text{ and } B = 0.0635$$

Hence

$$a = \text{antilog of } A = 10^{2.0478} = 111.6349$$

$$b = \text{antilog of } B = 10^{0.0635} = 1.1574$$

Hence the fitted equation is $y = 111.6349(1.1574)^x$. The estimated values of y (denoted by $\hat{y}$) are shown in the last column of Table E7.13(a).

Example E7.16

For the data given in Table E7.15, find the equation to best fitting curve of the form $xy^a = b$.

Table E7.16

x	200	150	100	60	40	10
y	1	1.5	1.8	2.4	4.1	6.5

Solution:

See Table E7.16(a).

Table E7.16(a)

	x	y	log x	log y	$(\log x)^2$	(log x)(log y)	$\hat{y}$ (estimated)
	200	1	2.3010	0	5.2947	0	1.1762
	150	1.5	2.1761	0.1761	4.7354	0.3832	1.4040
	100	1.8	2.0000	0.2553	4.0000	0.5105	1.8019
	60	2.4	1.7782	0.3802	3.1618	0.6761	2.4675
	40	4.1	1.6021	0.6128	2.5666	0.9817	3.1668
	10	6.5	1.0000	0.8129	1.0000	0.8129	7.4322
Σ	560	17.3	10.8573	2.2373	20.7585	3.3644	17.4485

Given

$$xy^a = b \tag{E.1}$$

Taking logarithms (to the base 10) on both sides of the above Equation (E.1), we get

$$\log x + a \log y = \log b \tag{E.2}$$

or

$$\frac{1}{a}\log x + \log y - \frac{\log b}{a} \tag{E.3}$$

The normal equations are given by

$$\Sigma Y = 6A + B\Sigma X \tag{E.4}$$

$$\Sigma XY = A\Sigma X + B\Sigma X^2 \tag{E.5}$$

where $Y = \log y,\ X = \log x, A = \frac{1}{a} \log b$ and $B = -1/a$.

Solving Equations (E.4) and (E.5), we obtain

$$A = 1.4865 \text{ and } B = -0.6154$$

Therefore $a = -1/B = -1/-0.6154 = 1.6250$

and $b = \text{antilog of } (aA) = 10^{(1.6250)(1.4865)} = 260.3529$

Hence, the fitted equation is $xy^{1.6250} = 260.3529$.

Example E7.17

Fit the following data:

x	0	2	4	6
y	11	16	19	26

to a straight line by considering that the data (2, 16) and (4, 19) are more significant or reliable with weights 6 and 11 respectively.

SOLUTION:

Weighted Least Squares Method

Let the straight line be $y = a + bx$. The normal equations are

$$a\Sigma w_i + b\Sigma w_i x_i = \Sigma w_i y_i \tag{E.1}$$

and

$$a\Sigma w_i x_i + b\Sigma w_i = \Sigma w_i x_i y_i \tag{E.2}$$

The values in Equations (E.1) and (E.2) are calculated as shown in Table E7.17.

Table E7.17

x	y	w	wx	wx^2	wy	wxy
0	11	1	0	0	11	0
2	16	6	12	24	96	192
4	19	11	44	176	209	836
6	26	1	6	36	26	156
Total		19	62	236	342	1184

The normal equations are

$$19a + 62b = 342 \tag{E.3}$$

and

$$62a + 236b = 1184 \tag{E.4}$$

The solution of Equations (E.3) and (E.4) gives

$$a = 11.4125 \text{ and } b = 2.0188$$

Hence, $$y = 11.4125 + 2.0188x$$

Estimation of error

x	y	w	Predicted y	Absolute error	(Absolute error)^2
0	11	1	11.4125	0.4125	0.1702
2	16	6	15.4500	0.5500	0.3025
4	19	11	19.4875	0.4875	0.2377
6	26	1	23.5250	2.4750	6.1256
Sum of squares of errors					6.8329

Example E7.18

Consider the Example E7.17 with the modified weights 300 and 50 instead of 6 and 11.

SOLUTION:

The modified calculations are shown in Table E7.18

Table E7.18

x	y	w	wx	wx^2	wy	wxy
0	11	1	0	0	11	0
2	16	30	60	120	480	960
4	19	50	200	800	950	3800
6	26	1	6	36	26	156
Total		82	266	956	1467	4916

The normal equations are

$$82a + 266b = 1467 \tag{E.1}$$

and $$266a + 956b = 4916 \tag{E.2}$$

The solution of Equations (E.1) and (E.2) gives

$$a = 12.4144 \text{ and } b = 1.6881$$

Hence, $$y = 12.4144 + 1.6881x$$

Estimation of error

x	y	w	Predicted y	Absolute error	(Absolute error)^2
0	11	1	12.4144	1.4144	2.0004
2	16	30	15.7905	0.2096	0.0439
4	19	50	19.1666	0.1666	0.0277
6	26	1	22.5427	3.4573	11.9530
Sum of squares of errors					14.0250

It is noted that when the weights on $x = 2$ and $x = 4$ are increased then the absolute error in y are reduced at these points, but, the sum of squares of errors is increased due to the less importance of the data (0, 11) and (6, 26).

7.24 SUMMARY

In this chapter, we have reviewed the relationship between two variables in two ways: (1) by using the regression analysis and (2) by computing the correlation coefficient. It was shown that the regression model can be used to evaluate the magnitude of change in one variable due to a certain change in another variable. The regression model also helps to predict the value of one variable for a given value of another variable. The correlation coefficient shows how strongly two variables are related. It does not, however, provide any information about the size of change in one variable as a result of a certain change in the other variable.

We have also reviewed the inferential methods in regression and correlation.

PROBLEMS

P7.1 Table P7.1 gives information on the monthly incomes (in hundreds of dollars) and monthly telephone bills (in dollars) for a random sample of 10 households.

Table P7.1

Income	16	45	35	31	30	14	40	15	36	40
Telephone bill	36	140	171	70	94	25	159	41	78	98

Use least-squares regression to determine the coefficients a and b in the function $y = a + bx$ that best fits the data.

P7.2 The following table P7.2 lists the annual incomes (in thousands of dollars) and amounts of life insurance (in thousands of dollars) of life insurance policies for six persons:

Table P7.2

Annual income	47	54	26	38	62	20
Life insurance	250	300	100	150	500	75

(*a*) find the regression line $y = a + bx$ with annual income as an independent variable and amount of life insurance policy as a dependent variable.

(*b*) determine the estimated value of life insurance of a person with an annual income of \$50,000.

P7.3 Find the least-squares regression line for the data on annual incomes and food expenditures of seven households given in Table P7.3. Use income as an independent variable and food expenditure as a dependent variable. All data is given in thousands of dollars.

Table P7.3

Income: x	35	50	22	40	16	30	25
Expenditure: y	9	15	6	11	5	8	9

P7.4 Table P7.4 gives data on age and crown-rump length for the fetuses. Use least-squares regression to determine the coefficients a and b in the function $y = a + bx$ that best fits the data:

Table P7.4

x	10	10	13	13	18	19	19	23	25	28
y	66	66	108	106	160	165	176	227	234	279

P7.5 The following data in Table P7.5 refers to the number of hours that 10 students studied for a math test and their scores on the test:

Table P7.5

Hours studied	1	17	22	12	7	4	14	10	9	4
Test score	21	83	90	60	45	38	74	66	59	32

(a) find the equation of the least-squares line that approximates the regression of the test scores on the number of hours studied.

(b) determine the average test score of a person who studied 15 hours for the test.

P7.6 The following Table P7.6 shows the first two grades, denoted by x and y respectively, of 10 students on two mid-term examinations in applied statistics.

(a) find the least squares regression line of y on x.

Table P7.6

Grade on first mid-term examination (x)	60	50	80	80	70	60	100	40	90	70
Grade on second mid-term examination (y)	80	70	70	90	50	80	95	60	80	60

P7.7 The following Table P7.7 shows ages x and systolic blood pressure y of 12 men.

(a) determine the least squares regression equation of y on x

(b) estimate the blood pressure of a man whose age is 45 years.

Table P7.7

Age (x)	56	42	72	36	63	47	55	49	38	42	68	60
Blood pressure (y)	147	125	160	118	149	128	150	145	115	140	152	155

P7.8 Table P7.8 shows the respective weight x and y of a sample of 12 fathers and their oldest sons.

(a) find the least-squares regression line of y on x.

Table P7.8

Weight of father, x (kg)	65	63	67	64	68	62	70	66	68	67	69	71
Weight of son, y (kg)	68	66	68	65	69	66	67	65	70	67	68	70

P7.9 Find the least squares regression line for the data on annual incomes and food-expenditures of seven households given in Table P7.9. Use income as independent variable and food expenditure as a dependent variable. The income and food-expenditures are in thousands of rupees.

Table P7.9

Income x	35	49	21	29	15	28	25
Food expenditure y	9	15	7	10	5	8	8.5

P7.10 A car manufacturing company wanted to investigate how the price of one of its car models depreciates with age. The company took a sample of eight cars of this model and collected the following information on the ages (in years) and prices (in hundreds of dollars) of these cars as shown in Table P7.10.

Table P7.10

Age	8	3	6	9	2	5	6	3
Price	16	74	40	19	120	36	33	86

(a) find the regression line $\hat{y} = a + bx$ with price as a dependent variable and age as independent variable

(b) give a brief interpretation of the values of a and b calculated in part *(a)*

(c) predict the price of a 7-year old car of this model

(d) estimate the price of an 4-year old car of this model.

For problems P7.11 to P7.20 do the following:

Fit a least squares regression line of the form $\hat{y} = a + bx$ for the data given in Tables P7.1 to P7.10 respectively. Assume x as the independent variable and y as the dependent variable.

(a) give a brief interpretation of the values of a and b calculated in $\hat{y} = a + bx$.

(b) compute the standard deviation of the sample errors, s_e.

(c) compute the error sum of squares, SSE.

(d) compute the total sum of squares, SST.

(e) compute the regression sum of squares, SSR.

(f) compute the coefficient of determination, r^2.

(g) compute the correlation coefficient, r.

P7.11 For the data given in Table P7.1.

P7.12 For the data given in Table P7.2.

P7.13 For the data given in Table P7.3.

P7.14 For the data given in Table P7.4.

P7.15 For the data given in Table P7.5.

P7.16 For the data given in Table P7.6.

P7.17 For the data given in Table P7.7.

P7.18 For the data given in Table P7.8.

P7.19 For the data given in Table P7.9.

P7.20 For the data given in Table P7.10.

P7.21 Construct a 95% confidence interval for B for the data in Problem P7.1.

P7.22 Construct a 95% confidence interval for B for the data in Problem P7.2.

P7.23 Construct a 95% confidence interval for B for the data in Problem P7.3.

P7.24 Construct a 95% confidence interval for B for the data in Problem P7.4.

P7.25 Construct a 95% confidence interval for B for the data in Problem P7.5.

P7.26 Construct a 95% confidence interval for B for the data in Problem P7.6.

P7.27 Construct a 95% confidence interval for B for the data in Problem P7.7.

P7.28 Construct a 95% confidence interval for B for the data in Problem P7.8.

P7.29 Construct a 95% confidence interval for B for the data in Problem P7.9.

P7.30 Construct a 95% confidence interval for B for the data in Problem P7.10.

P7.31 Test at the 1% significance level if the slope of the regression line in Problem P7.1 is positive.

P7.32 Test at the 1% significance level if the slope of the regression line in Problem P7.2 is positive.

P7.33 Test at the 1% significance level if the slope of the regression line in Problem P7.3 is positive.

P7.34 Test at the 1% significance level if the slope of the regression line in Problem P7.4 is positive.

P7.35 Test at the 1% significance level if the slope of the regression line in Problem P7.5 is positive.

P7.36 Test at the 1% significance level if the slope of the regression line in Problem P7.6 is positive.

P7.37 Test at the 1% significance level if the slope of the regression line in Problem P7.7 is positive.

P7.38 Test at the 1% significance level if the slope of the regression line in Problem P7.8 is positive.

P7.39 Test at the 1% significance level if the slope of the regression line in Problem P7.9 is positive.

P7.40 Test at the 1% significance level if the slope of the regression line in Problem P7.10 is positive.

P7.41 Refer to Problem P7.1

(a) construct a 90% confidence level for B

(b) test at the 5% significance level if B is negative.

P7.42 Refer to Problem P7.2

(a) construct a 90% confidence level for B

(b) test at the 5% significance level if B is negative.

P7.43 Refer to Problem P7.3

(a) construct a 90% confidence level for B

(b) test at the 5% significance level if B is negative.

P7.44 Refer to Problem P7.4

(a) construct a 90% confidence level for B

(b) test at the 5% significance level if B is negative.

P7.45 Refer to Problem P7.5

(a) construct a 90% confidence level for B

(b) test at the 5% significance level if B is negative.

P7.46 Refer to Problem P7.6

(a) construct a 90% confidence level for B

(b) test at the 5% significance level if B is negative.

P7.47 Refer to Problem P7.7

(a) construct a 90% confidence level for B

(b) test at the 5% significance level if B is negative.

P7.48 Refer to Problem P7.8

(*a*) construct a 90% confidence level for B

(*b*) test at the 5% significance level if B is negative.

P7.49 Refer to Problem P7.9

(*a*) construct a 90% confidence level for B

(*b*) test at the 5% significance level if B is negative.

P7.50 Refer to Problem P7.10

(*a*) construct a 90% confidence level for B

(*b*) test at the 5% significance level if B is negative.

P7.51 Refer to Problem P7.1. Find a 99% confidence interval for the mean value of telephone bill, y, when the income, $x = \$4200$.

P7.52 Refer to Problem P7.2. Find a 99% confidence interval for the mean value of life insurance, y, when the annual income, $x = \$53000$.

P7.53 Refer to Problem P7.3. Find a 99% confidence interval for the mean value of expenditure, y, when the income, $x = \$42000$.

P7.54 Refer to Problem P7.4. Find a 99% confidence interval for the mean value of y, when the value of $x = 20$.

P7.55 Refer to Problem P7.5. Find a 99% confidence interval for the mean value of test score, y, when the hours studied, $x = 15$.

P7.56 Refer to Problem P7.6. Find a 99% confidence interval for the mean value of grade on second mid-term examination, y, when the grade on first mid-term examination, $x = 82$.

P7.57 Refer to Problem P7.7. Find a 99% confidence interval for the mean value of blood pressure, y, when the age, $x = 65$.

P7.58 Refer to Problem P7.8. Find a 99% confidence interval for the mean value of weight of son, y, when the weight of father, $x = 72$.

P7.59 Refer to Problem P7.9. Find a 99% confidence interval for the food expenditure, y, when the income, $x =$ Rs.41,000.

P7.60 Refer to Problem P7.10. Find a 99% confidence interval for the price, y, when the age of the car, $x = 7$ years.

P7.61 Fit $y = bx^m$ (power function) in Problem P7.1 using a logarithmic transformation.

P7.62 Fit $y = bx^m$ (power function) to the data in Problem P7.2 using a logarithmic transformation.

P7.63 Fit $y = bx^m$ (power function) to the data in Problem P7.3 using a logarithmic transformation.

P7.64 Fit $y = bx^m$ (power function) to the data in Problem P7.4 using a logarithmic transformation.

P7.65 Fit $y = bx^m$ (power function) to the data in Problem P7.5 using a logarithmic transformation.

P7.66 Determine the coefficient of the polynomial $y = a + bx + cx^2$ that best fit the data given in the following table.

x	1	3	5	7	10
y	2.1	5.1	5.45	6.12	6.62

Determine the standard error of the estimate and correlation coefficient.

P7.67 The following data were collected in an experiment to study the relationship between shear strength in kPa (y) and curing temperature in °C (x).

x	1.38	1.40	1.46	1.48	1.52	1.53
y	5.392	5.612	5.671	5.142	4.481	4.129

(a) fit a least-squares quadratic model of the form $y = a + bx + cx^2$ to the above data

(b) using the equation, compute the residuals.

(c) compute the error sum of squares and total sum of squares.

(d) compute the error variance estimate.

(e) compute the coefficient of determination.

P7.68 The following data were collected in an experiment to study the relationship between the number of kilograms of fertilizer (x) and the yield of tomatoes in bushels (y).

x	5	10	30	40	50
y	32	42	54	50	42

(a) fit a least-squares quadratic model of the form $y = a + bx + cx^2$ to the above data.

(b) using this equation, compute the regression sum of squares $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$.

(c) compute the error sum of squares $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and total sum of squares $\sum_{i=1}^{n}(y_i - \bar{y})^2$.

(d) compute the error variance estimate *(b)* + *(c)*.

(e) compute the coefficient of determination, r^2.

P7.69 Fit a least-square parabola $y = a + bx + c^2$ to the following data:

x	0	1	2	3	4	5	6
y	2.4	2.1	3.2	5.6	9.3	14.6	21.9

Determine the coefficient of determination.

P7.70 The following table gives the data collected in an experiment to study the relationship between the stopping distance $d(m)$_ of an automobile traveling at speeds v(km/hr) at the instant the danger is sighted.

(a) fit a least-squares parabola of the form $d = a + bv + cv^2$ to the data

(b) determine the coefficient of determination.

Speed v(km/hr)	32	48	64	80	96	112
Stopping distance d(m)	16.5	27.5	19.5	24.5	29.3	34.2

P7.71 For the data given in Problem P7.1, find a 99% prediction interval for y for a randomly selected value of $x = 42$.

P7.72 For the data given in Problem P7.2, find a 99% prediction interval for y for a randomly selected value of $x = 53$.

P7.73 For the data given in Problem P7.3, find a 99% prediction interval for y for a randomly selected value of $x = 42$.

P7.74 For the data given in Problem P7.4, find a 99% prediction interval for y for a randomly selected value of $x = 20$.

P7.75 For the data given in Problem P7.5, find a 99% prediction interval for y for a randomly selected value of $x = 15$.

P7.76 For the data given in Problem P7.6, find a 99% prediction interval for y for a randomly selected value of $x = 82$.

P7.77 For the data given in Problem P7.7, find a 99% prediction interval for y for a randomly selected value of $x = 45$.

P7.78 For the data given in Problem P7.8, find a 99% prediction interval for y for a randomly selected value of $x = 72$.

P7.79 For the data given in Problem P7.9, find a 99% prediction interval for y for a randomly selected value of $x = 41$.

P7.80 For the data given in Problem P7.10, find a 99% prediction interval for y for a randomly selected value of $x = 7$.

P7.81 Use multiple linear regression fit of the form $y = a + bx_1 + cx_2$ for the following data:

x_1	0	1	1	2	2	3	3	4	4
x_2	0	1	2	1	2	1	2	1	2
y	15	18	12.8	25.7	20.4	35	30	45.3	40.1

Compute the coefficients, the standard error of the estimate, and the correlation coefficient.

P7.82 Use multiple linear regression fit of the form $y = a + bx_1 + cx_2$ for the following data:

x_1	0	0	1	2	1	1.5	3	3	–1
x_2	0	1	0	1	2	1	2	3	–1
y	1	6	4	–4	–2	–1.5	–12	–15	17

Compute the coefficients, the standard error of estimate, and the correlation coefficient.

P7.83 Use multiple linear regression fit of the form $y = a + bx_1 + cx_2$ for the following data:

x_1	0	0	1	1	2	3	0	2	1	4
x_2	0	1	0	1	2	0.5	2	3	4	1
y	3	8	7	12	21	15	13	26	27	24

Compute the coefficients, the standard error of estimate, and the correlation coefficient.

P7.84 Use multiple linear regression fit of the form $y = a + bx_1 + cx_2$ for the following data:

x_1	0	0	1	1	2	0	1	2	1	1
x_2	0	1	0	1	0	2	2	1	3	1
y	23	15	19	10	15	5	0	5	–10	0

Compute the coefficients, the standard error of estimate, and the correlation coefficient.

P7.85 Use multiple linear regression fit of the form $y = a + bx_1 + cx_2$ for the following data:

x_1	0	0	1	1	2	0	1	2	1	3
x_2	0	1	0	1	0	2	2	1	3	1
y	29	10	23	4	19	–10	–16	–2	–36	–8

Compute the coefficients, the standard error of estimate, and the correlation coefficient.

P7.86 For the data given in Table P7.86, find the equation to the best fitting exponential curve of the form $y = ae^{bx}$.

Table P7.86

x	1	2	3	4	5
y	100	90	80	75	70

P7.87 For the data given in Table P7.87, find the equation to the best fitting exponential curve of the form $y = ae^{bx}$.

Table P7.87

x	2	3	4	5	6
y	3.8	5.8	7.8	8.8	9.8

P7.88 For the data given in Table P7.88, find the equation to the best fitting exponential curve of the form $y = ae^{bx}$.

Table P7.88

x	2.2	3	4	6	7
y	31	38	45	68	84

P7.89 For the data given in Table P7.89, find the equation to the best fitting exponential curve of the form $y = ab^x$.

Table P7.89

x	1	2	3	4	5
y	22	8	3	1	0.35

P7.90 For the data given in Table P7.90, find the equation to the best fitting exponential curve of the form $y = ab^x$.

Table P7.90

x	2	4	6	8	10
y	3	13	32	57	91

P7.91 For the data given in Table P7.91, find the equation to the best fitting exponential curve of the form $y = ab^x$.

Table P7.91

x	1	3	5	7	9
y	3	2	1.3	0.72	0.43

P7.92 For the data given in Table P7.92, find the equation to the best fitting exponential curve of the form $y = xy^a = b$.

Table P7.92

x	190	134	89	55	37	8.9
y	0.97	1.14	1.32	1.63	1.92	3.5

P7.93 For the data given in Table P7.93, find the equation to the best fitting exponential curve of the form $y = xy^a = b$.

Table P7.93

x	2	3	5	7	9	11
y	1.25	1.21	1.16	1.14	1.11	1.10

P7.94 For the data given in Table P7.94, find the equation to the best fitting exponential curve of the form $y = xy^a = b$.

Table P7.94

x	232	178	99	66	51
y	1.1	1.3	1.8	2.2	2.5

P7.95 Find a non-linear relationship of the form $y = a + b \log x$ for the data given in Table P7.95. Determine the linear correlation coefficient.

Table P7.95

x	1.2	4.7	8.3	20.9
y	0.6	5.1	6.9	10

P7.96 Fit the following data to a straight line $y = a + bx$ by considering the weights as given in the table. Compute the sum of squares of errors.

x	1	17	22	12	7	4	14	10	9	4
y	21	83	90	60	45	38	74	66	59	32
w	5	1	7	1	1	8	1	11	1	4

P7.97 Fit the following data to a straight line $y = a + bx$ by considering the weights as given in the table. Compute the sum of squares of errors.

x	16	45	35	31	30	14	40	15	36	40
y	50	134	107	95	90	44	120	47	110	120
w	1	6	1	3	1	7	1	11	1	16

P7.98 Fit the following data to a straight line $y = a + bx$ by considering the weights as given in the table. Compute the sum of squares of errors.

x	47	54	26	38	62	20
y	250	300	100	150	500	75
w	1	5	1	3	1	7

P7.99 Fit the following data to a straight line $y = a + bx$ by considering the weights given in the table. Compute the sum of squares of errors.

x	35	50	22	40	16	30	25
y	9	15	6	11	5	8	9
w	1	2	1	2	1	3	4

P7.100 Fit the following data to a straight line $y = a + bx$ by considering the weights given in the table. Compute the sum of squares of errors.

x	10	10	13	13	18	19	19	23	25	28
y	66	66	108	106	160	165	176	227	234	279
w	2	1	3	1	4	1	5	1	6	1

P7.101 Fit the following data to a straight line $y = a + bx$ by considering the weights given in the table. Compute the sum of squares of errors.

x	1	17	22	12	7	4	14	10	9	4
y	21	83	90	60	45	38	74	66	59	32
w	1	2	1	3	1	2	1	4	1	5

P7.102 Fit the following data to a straight line $y = a + bx$ by considering the weights given in the table. Compute the sum of squares of errors.

x	1	17	22	12	7	4
y	21	83	90	60	45	38
w	5	1	7	1	1	8

P7.103 Fit the following data to a straight line $y = a + bx$ by considering the weights given in the table. Compute the sum of squares of errors.

x	16	45	35	31	30
y	50	134	107	95	90
w	1	6	1	3	1

P7.104 Fit the following data to a straight line $y = a + bx$ by considering the weights given in the table. Compute the sum of squares of errors.

x	47	54	26	38	62	20
y	250	300	100	150	500	75
w	1	4	1	2	1	7

P7.105 Fit the following data to a straight line $y = a + bx$ by considering the weights given in the table. Compute the sum of squares of errors.

x	35	50	22	40
y	9	15	6	11
w	2	3	4	2

P7.106 Fit the following data to a straight line $y = a + bx$ by considering the weights given in the table. Compute the sum of squares of errors.

x	10	10	13	13	18
y	66	66	108	106	160
w	5	2	3	7	4

P7.107 Fit the following data to a straight line $y = a + bx$ by considering the weights given in the table. Compute the sum of squares of errors.

x	1	17	22	12	7
y	21	75	96	60	45
w	1	2	1	3	1

P7.108 Construct a least-squares quadratic approximation to the function $y = e^x$ on $[0, 1]$.

P7.109 Construct a least-squares quadratic approximation to the function $y = x \ln x$ on $[1, 3]$.

P7.110 Construct a least-squares quadratic approximation to the function $y = x^3$ on $[0, 2]$.

P7.111 Construct a least-squares quadratic approximation to the function $y = \frac{1}{x}$ on $[1, 3]$.

P7.112 Construct a least-squares quadratic approximation to the function $y = x^2 + 3x + 2$ on $[0, 1]$.

P7.113 Use the Gram-Schmidt orthogonalization process to construct $\phi_0(x)$, $\phi_1(x)$, $\phi_2(x)$, and $\phi_3(x)$ for the interval $[0, 1]$.

P7.114 Use the Gram-Schmidt orthogonalization process to construct $\phi_0(x)$, $\phi_1(x)$, $\phi_2(x)$, and $\phi_3(x)$ for the interval $[0, 2]$.

P7.115 Use the Gram-Schmidt orthogonalization process to construct $\phi_0(x)$, $\phi_1(x)$, $\phi_2(x)$, and $\phi_3(x)$ for the interval $[1, 3]$.

REVIEW QUESTIONS

1. Explain the words *simple* and *linear* as used in simple *linear regression.*
2. What is meant by independent and dependent variables in a regression model?
3. What is the difference between exact and non-exact relationships between two variables?
4. What is the difference between linear and nonlinear relationships between two variables?
5. Explain the difference between a simple and multiple regression model.
6. Explain the difference between a deterministic and a probability regression model.
7. What is the reason for the inclusion of the random error term in a regression model?
8. Describe the least-squares method and least squares regression line.

9. Describe the meaning and concept of SSE.
10. What is the difference between y and $\hat{y}$?
11. If two variables x and y have a positive linear relationship, what happens to the value of y as x increases?
12. If two variables x and y have a negative linear relationship, what happens to the value of y when x increases?
13. Describe the following terms:
    - *(a)* population regression line
    - *(b)* true values of A and B
    - *(c)* sample regression line
    - *(d)* estimated values of A and B that are denoted by a and b, respectively.
14. Describe the assumptions made in a population regression model.
15. What are the degrees of freedom for a simple linear regression model?
16. Describe the meaning of *coefficient of determination.*
17. Explain the meaning of SST and SSR.
18. Describe the mean, standard deviation, and shape of the sampling distribution of the slope b of the simple linear regression model.
19. Explain the meaning of a linear correlation coefficient between two variables.
20. Describe the following terms:
    - *(a)* no linear correlation
    - *(b)* weak positive linear correlation
    - *(c)* weak negative linear correlation
    - *(d)* strong negative linear correlation
    - *(e)* strong positive linear correlation
    - *(f)* perfect positive linear correlation
    - *(g)* perfect negative linear correlation
21. What kind of plot is useful for deciding whether finding a regression line for a set of data points in reasonable?
22. Identify one use of a regression equation.
23. Regarding the variables in a regression analysis,
    - *(a)* what is the independent variable called?
    - *(b)* what is the dependent variable called?
24. In the context of regression analysis, what is an
    - *(a)* outlier?
    - *(b)* influential observation?
25. Identify the use of the coefficient of determination as a descriptive measure.
26. For each of the sum of squares in regression, state its name and what it measures.
    - *(a)* SST
    - *(b)* SSR
    - *(c)* SSE

STATE TRUE OR FALSE

1. The graph of a linear equation with one independent variable is a line. (True/False)
2. The graph of a linear equation slopes upward unless the slope is 0. (True/False)
3. The value of the y-intercept has no effect on the direction that the graph of a linear equation slopes. (True/False)
4. A simple regression is a regression model that contains both a and b. (True/False)
5. The relationship between independent and dependent variable represented by the simple linear regression is that of a *curve*. (True/False)
6. A deterministic regression model is a model that does not contain the random error term. (True/False)
7. A probabilistic regression model is a model that does not contain the random error term. (True/False)
8. The least-squares regression line minimizes the sum of errors. (True/False)
9. The degrees of freedom for a simple regression model are $n - 2$. (True/False)
10. The coefficient of determination gives the proportion of total squared errors (SST) that is explained by the use of the regression model. (True/False)
11. The linear coefficient measures the strength of the linear association between two variables. (True/False)
12. The value of the coefficient of determination is always in the range –1 to +1. (True/False)
13. The value of the correlation coefficient is always in the range –1 to +1. (True/False)
14. The symbol used for the linear correlation coefficient is r^2. (True/False)
15. A value of r close to ±1 indicates that there is a weak linear relationship between the variables. (True/False)
16. A value of r close to zero indicates that there is a either no linear relationship between the variables or a weak one. (True/False)
17. If y tends to increase linearly as x increases, the variables are positively linearly correlated. (True/False)
18. If y tends to decrease linearly as x increases, the variables are positively linearly correlated. (True/False)
19. If there is no linear relationship between x and y, the variables are linearly uncorrelated. (True/False)
20. If there is a very strong positive correlation between two variables, a casual relationship exists between the two variables. (True/False)
21. The independent variable in a linear relationship $y = a + bx$ is x. (True/False)
22. The dependent variable in a linear relationship $y = a + bx$ is x. (True/False)
23. The slope in a linear relationship $y = a + bx$ is b. (True/False)
24. The y-intercept of a linear relationship $y = a + bx$ is a. (True/False)
25. The y-intercept of a line has no effect on the steepness of the line. (True/False)
26. A horizontal line has no slope. (True/False)

27. If a line has a positive slope, y-values on the line decrease as the x-value decrease. (True/False)
28. Based on the least-squares criterion, the line that best fits a set of data points is the one having *smallest* possible sum of squared errors. (True/False)
29. The line that best fits a set of data points according to the least-squares criterion is called the *regression* line. (True/False)
30. Using a regression equation to make predictions for values of the predictor variable outside the range of the observed values of the predictor variable is called *extrapolation.* (True/False)
31. One use of the linear correlation coefficient is as a descriptive measure of the strength of the *linear* relationship between two variables. (True/False)
32. A positive linear relationship between two variables means that one variable tends to increase linearly as the other *decrease.* (True/False)
33. A value of r close to -1 suggests a strong *positive* linear relationship between the variables. (True/False)
34. A value of r close to 0 suggests at most a weak linear relationship between the variables. (True/False)
35. A strong correlation between two variables doesn't necessarily mean they are casually related. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **True** 4. **True** 5. **False** 6. **True** 7. **False** 8. **False** 9. **True** 10. **True**

11. **True** 12. **True** 13. **True** 14. **False** 15. **False** 16. **True** 17. **True** 18. **False** 19. **True** 20. **True**

21. **True** 22. **False** 23. **True** 24. **True** 25. **True** 26. **False** 27. **True** 28. **True** 29. **True** 30. **True**

31. **True** 32. **False** 33. **False** 34. **True** 35. **True**

❑ ❑ ❑

CHAPTER 8

chi-square Tests

8.1 INTRODUCTION

In Chapter 6, the tests of hypotheses about the mean, the difference between two means, and proportion, and the difference between two proportions were presented. In this chapter, we will describe three types of tests:

1. Tests of hypotheses for experiments with more than two categories, called goodness-of-fit tests.
2. Tests of hypotheses about contingency table, called independence and homogeneity tests.
3. Tests of hypotheses about the variance and standard deviation of a single population.

All these tests are performed by using *the chi-square distribution*. The description of chi-square distribution was presented in Chapter 4, Section 4.10.

8.2 A GOODNESS-OF-FIT TEST

In this section, we describe how to make tests of hypotheses about experiments with more than two possible outcomes or categories. These experiments are called *multinomial experiments* and possess certain characteristics as described below.

An experiment with the following characteristics is called a *multinomial experiment*.

(a) The experiment consists of n identical trials (repetitions).
(b) Each trial results in one of k possible outcomes or categories where $k > 2$.
(c) The trials are all independent.
(d) The probabilities of the various outcomes remain constant for each trial.

An experiment of many rolls of a die is an example of a multinomial experiment. The experiment consists of several identical rolls (trials) and each roll (trial) results in one of the six possible outcomes. Each roll is considered independent of the other rolls and the probabilities of the six outcomes remain constant for each roll. A binomial experiment is a special case of a multinomial experiment.

Observed and Expected Frequencies

The frequencies obtained from the performance of an experiment are called the *observed frequencies* and denoted by O. The *expected frequencies*, denoted by E, are the frequencies that we will expect to obtain if the null hypothesis is true. The expected frequency for a category is obtained as

$$E = np \tag{8.1}$$

where n = the sample size

p = the probability that an element belongs to that category if the null hypothesis is true.

Degrees of Freedom for a Goodness-of-Fit Test

In a goodness-of-fit test, the degrees of freedom are given by

$$df = k - 1 \tag{8.2}$$

where k = the number of possible outcomes or categories for an experiment.

Note that the *chi-square goodness-of-fit test is always a right tailed test.*

The Statistic for a Goodness-of-Fit Test

The test statistic for a goodness-of-fit test is χ^2 and its value is computed from

$$\chi^2 = \sum \frac{(O - E)^2}{E} \tag{8.3}$$

where O = observed frequency for a category

E = expected frequency for a category = np.

To make a goodness-of-fit test, the sample size must be large enough so that the expected frequency for each category is at least 5. Table 8.1 gives a step-by-step method for conducting a chi-square goodness-of-fit test.

Table 8.1: chi-square goodness-of-fit test

Purpose: To perform a hypothesis test for the distribution of a variable

Assumptions:

1. All expected frequencies are 1 or greater
2. At most 20% of the expected frequencies are less than 5
3. Simple random sample

Step 1: The null and alternative hypotheses are

H_0: The variable has the specified distribution.

H_1: The variable does not have the specified distribution.

Step 2: Calculate the expected frequency for each possible value of the variable by using the formula $E = np$, where n is the sample size and p is the relative frequency (or probability) given for the value in the null hypothesis.

Step 3: Determine whether the expected frequencies satisfy Assumptions 1 and 2. If they do not, this procedure should not be used.

Step 4: Decide on the significance level, α

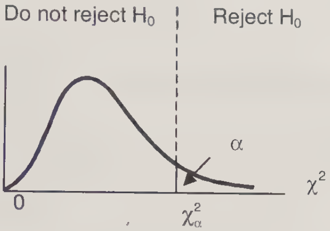
Step 5: Compute the value of the test statistic

$$\chi^2 = \Sigma(O - E)^2/E$$

where O and E represent observed and expected frequencies, respectively. Denote the value of the test statistic χ^2.

Critical-value Approach

Step 6: The critical value is χ^2_α with $df = k - 1$, where k is the number of possible values for the variable. Use table in Appendix-*F* to find the critical value.



Step 7: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

Step 8: Interpret the results of the hypothesis test.

Example E8.1

A home mail order company likes to determine if the number of orders received on each day of the five working days of the week is the same. Table E8.1 lists the frequency distribution for these orders received from a sample of 400 orders during a four-week period.

Table E8.1

Day of the week	Monday	Tuesday	Wednesday	Thursday	Friday
Number of orders received	90	69	67	88	86

Test at the 5% significance level if the null hypothesis that the orders are evenly distributed overall days of the week is true.

SOLUTION:

H_0 : orders are evenly distributed overall days of the week.

H_1 : orders are not evenly distributed overall days of the week.

$$df = k - 1 = 5 - 1 = 4$$

For $\alpha = 0.05$ and $df = 4$, the critical value of $\chi^2 = 9.488$ [From the chi-square distribution table in Appendix-*F*].

Day	O	p	E = np	O – E	$(O - E)^2$	$(O - E)^2/E$
Monday	90	0.2	80	10	100	1.250
Tuesday	69	0.2	80	–11	121	1.513
Wednesday	67	0.2	80	–13	169	2.113
Thursday	88	0.2	80	8	64	0.800
Friday	86	0.2	80	6	36	0.450
Total	400	1.0	400	0	490	6.125

The values of the test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Since 6.125 < 9.488, do not reject H_0. Hence, we conclude that orders are evenly distributed over all days of the week.

Example E8.2

Table E8.2 presents the frequency distribution of automobiles sold at an automobile dealership during the past 12 months.

Table E8.2

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Automobiles sold	20	18	18	13	12	16	14	17	23	25	28	31

Can we reject the null hypothesis that the number of automobiles sold at this automobile dealership is the same for each month at the 10% significance level?

SOLUTION:

H_0 : The number of automobiles sold is the same for each month.

H_1 : The number of automobiles sold is not the same for each month.

$$df = k - 1 = 12 - 1 = 11$$

From the chi-square distribution table in Appendix-*F*, the critical value of χ^2 for $df = 11$ and 0.10 area in the right tail is 17.275 [From the chi-square distribution Table in Appendix-*F*].

The number of automobiles sold will be the same for each month if 1/12th of the automobiles sold during the full year are sold each month. Table E8.2(a) shows the computations.

Table E8.2(a)

Month	O	p	E = np	O – E	(O – E)^2	(O – E)^2/E
January	20	0.0833	19.5849	0.4151	0.1723	0.0088
February	18	0.0833	19.5849	–1.5849	2.5119	0.1283
March	18	0.0833	19.5849	–1.5849	2.5119	0.1283
April	13	0.0833	19.5849	–6.5849	43.3609	2.2140
May	12	0.0833	19.5849	–7.5849	57.5307	2.9375
June	16	0.0833	19.5849	–3.5849	12.8515	0.6562
July	14	0.0833	19.5849	–5.5849	31.1911	1.5926
August	17	0.0833	19.5849	–2.5849	6.6847	0.3412
September	23	0.0833	19.5849	3.4151	11.6629	0.5955
October	25	0.0833	19.5849	5.4151	29.3233	1.4972
November	28	0.0833	19.5849	8.4151	70.8139	3.6157
December	31	0.0833	19.5849	11.4151	130.3045	6.6533
Total	235	1.0	235.0	0	398.9167	20.3686

Hence, the value of the test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 20.3686$$

Reject the null hypothesis. The number of automobiles sold is not the same for each month.

8.3 CONTINGENCY TABLE

When information on more than one variable for each element is summarised and presented in a two-way classification table, such a table is called *contingency table* or *cross-tabulation*. Table 8.2 is an example of a contingency table which contains the results obtained for four samples of fixed sizes selected from four populations.

Table 8.2

	Sample selected from			
	Population 1	Population 2	Population 3	Population 4
Row 1	25	80	60	123
Row 2	47	65	90	73
Row 3	19	40	106	94

A contingency table can be of any size. In general, an $R \times C$ table contains R rows and C columns. Each of the twelve boxes that contain the numbers in Table 8.2 is called a *cell*. In Table 8.2, the number of cells = $R \times C = 3 \times 4 = 12$ cells. The numbers written inside the cells are generally called the *joint frequencies*.

8.4 A TEST OF INDEPENDENCE OR HOMOGENEITY

The tests of independence or homogeneity are performed using the contingency table described in section 8.3. The procedure to make the tests is very similar to the one applied in for a goodness-of-fit tests.

A test of independence involves a test of the null hypothesis that two attributes of a population are not related. The *degrees of freedom for a test of independence* are given by

$$df = (R-1)(C-1) \tag{8.4}$$

where R and C are the number of rows and the number of columns, respectively, in a given contingency table.

Test Statistic for a Test of Independence:

The value of the *test statistic* χ^2 *for a test of independence* is obtained from

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{8.5}$$

where O and E are the observed and expected frequencies, respectively, for a cell.

The null hypothesis in a test of independence is always that the two attributes are not related. The alternative hypothesis is that the two attributes are related. The frequencies obtained from the performance of an experiment for a contingency table are called the *observed frequencies*. The procedure to calculate the *expected frequencies* for a contingency table for a test of independence is slightly different from the one for a goodness-of-fit test.

Expected Frequencies for a Test of Independence

The expected frequency E for a cell is calculated from $E = \dfrac{\text{(row total) (column total)}}{\text{sample size}}$

A step-by-step method for conducting a chi-square independence test by using the critical value approach is given in Table 8.3.

Table 8.3: chi-square independence test

Purpose: To perform a hypothesis test to decide whether two variables are associated

Assumptions:

1. All expected frequencies are 1 or greater
2. At most 20% of the expected frequencies are less than 5
3. Simple random sample

Step 1: The null and alternative hypotheses are

H_0: The two variables are not associated.

H_1: The two variables are associated.

Step 2: Calculate the expected frequencies by using the formula $E = RC/n$, where R = row total, C = column total, and n = sample size. Place each expected frequency below its corresponding observed frequency in the contingency table.

Step 3: Determine whether the expected frequencies satisfy Assumptions 1 and 2. If they do not, this procedure should not be used.

Step 4: Decide on the significance level,α

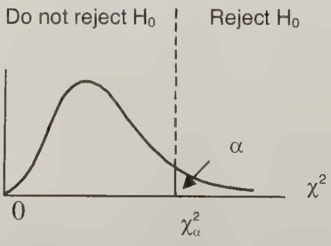
Step 5: Compute the value of the test statistic

$$\chi^2 = \Sigma(O - E)^2/E$$

where O and E represent observed and expected frequencies, respectively. Denote the value of the test statistic χ^2.

Critical-value Approach

Step 6: The critical value is χ^2_α with $df = (R - 1) \times (C - 1)$, where R and C are the number of possible values for the two variables. Use table in Appendix-F to find the critical value.



Step 7: If the value of the test statistic falls in the rejection region, reject H_0; otherwise, do not reject H_0.

Step 8: Interpret the results of the hypothesis test.

Example E8.3

Table E8.3 shows the results of a random survey of 100 people who were checked by police to see if they were wearing seat belts. Test at the 2.5% significance level if being a male or a female and wearing or not wearing a seat belt are related.

Table E8.3

	Wearing seat belt	Not wearing seat belt
Male	35	20
Female	30	15

SOLUTION:

H_0 : Gender and wearing or not wearing of seat belt are not related.

H_1 : Gender and wearing or not wearing of seat belt are related.

$$df = (R-1)(C-1) = (2-1)(2-1) = 1$$

Table E8.3(a)

	Wearing a seat belt	Not wearing a seat belt	Total
Men	35 (35.75)	20 (19.25)	55
Women	30 (29.25)	15 (15.75)	45
Total	65	35	100

The numbers in parentheses are expected frequencies, which are calculated using the formula

$$E = \frac{\text{(row total) (column total)}}{\text{total of both samples}}$$

The values of the test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 0.0157 + 0.0292 + 0.0192 + 0.0357 = 0.0999$$

Since 0.0999 < 5.024 [From the chi-square distribution table in Appendix-*F*], do not reject the null hypothesis. Being a male or female and wearing or not wearing a seat belt are *not* related.

Example E8.4

A study conducted to find the relationship between smoking and household income (Rupees per month) for males aged 15 years and older in a large city in India gave the two-way distribution of the observed frequencies based on household incomes and smoking status as shown in Table E8.4.

Table E8.4

Household income (in Rupees)	Current smokers	Former smokers	Never smokers
Less than 5,000	360	20	190
5,000 to 6,999	500	15	200
7,000 to 9,999	650	16	280
10,000 to 14,999	460	10	270
15,000 or higher	300	12	150

Using the 5% significance level, can we calculate that household incomes and smoking status are dependent for people living in that city?

SOLUTION:

H_0 : Household income and smoking status are independent.

H_1 : Household income and smoking status are dependent.

$$df = (R-1)(C-1) = (5-1)(3-1) = 8$$

From the chi-square distribution table in Appendix-*F*, the critical value of χ^2 for $df = 8$ and 0.05 area in the right tail is 15.507.

The value of the test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 0.758 + 5.122 + 0.45 + 1.567 + 0.003 + 3.215 + 0.958 + 0.842 + 1.38 + 1.756 + 2.091 + 5.227 + 0.099 + 0.482 + 0.075 = 23.369$$

Reject the null hypothesis. Household income and smoking status are dependent for people living in that city.

Table E8.4(a) shows the detailed computations. The numbers in parentheses are expected frequencies, which are calculated using the formula $E = \frac{\text{(row total) (column total)}}{\text{total of both samples}}$.

Table E8.4(a)

Household income	Current smokers	Former smokers	Never smokers	Total
< 5000	360 (376.901)	20 (12.121)	190 (180.979)	570
5000–6999	500 (422.779)	15 (15.204)	200 (227.017)	715
7000–9999	650 (625.523)	16 (20.116)	280 (300.361)	946
10000–14999	460 (489.31)	10 (15.736)	270 (234.955)	740
≥ 15000	300 (305.488)	12 (9.824)	150 (146.688)	462
Total	2270	73	1090	3433

Example E8.5

Two brands of drugs were experimented on two groups of randomly selected patients for a particular illness. Table E8.5 shows the information about the number of patients who cured and not cured by the administration of the two brands of drugs.

Table E8.5

	Cured	Not cured
Brand A	45	15
Brand B	17	23

Test at the 1% significance whether or not the two brands are similar in curing and not curing the patients.

SOLUTION:

H_0 : The two brands are similar in curing the patients.

H_1 : The two brands are *not* similar in curing the patients.

$$df = (R-1)(C-1) = (2-1)(2-1) = 1$$

From the chi-square distribution table in Appendix-*F*, the critical value of χ^2 for $df = 1$ and 0.01 area in the right tail is 6.635.

Table E8.5(a)

	Cured	Not cured	Total
Brand A	45 (37.2)	15 (22.8)	60
Brand B	17 (24.8)	23 (15.2)	40
Total	62	38	100

The numbers in parentheses are expected frequencies, which are calculated using the formula:

$$E = \frac{\text{(row total) (column total)}}{\text{(total of both samples)}}$$

The value of the test statistic is $\chi^2 = \sum \frac{(O-E)^2}{E} = 1.6355 + 2.6684 + 2.4532 + 4.0026 = 10.7598$

Reject the null hypothesis. The two drug brands are not similar in curing the patients.

A Test of Homogeneity

A *test of homogeneity* involves testing the null hypothesis that the proportions of elements with certain characteristics in two or more different populations are the same against the alternative hypothesis that these proportions are not the same.

The procedure to make a test of homogeneity is very similar to the procedure used to make a test of independence discussed in earlier section. A test of homogeneity is also right tailed like the test of independence.

Example E8.6

A survey of 500 adults (250 men and 250 women) conducted in a particular city in India gave their opinions on weather forecasts and the greenhouse effect as given in the two-way classification table below:

Table E8.6

	Men	Women
Have been big changes	85	115
Have not been	110	70
Only small changes	40	30
Do not know (no opinion)	15	35

Is there sufficient evidence from the data to reject the null hypothesis that men and women have similar opinions about the greenhouse effect at 1% significance level?

SOLUTION:

H_0 : Men and women have similar opinions about the greenhouse effect.

H_1 : Men and women *do not* have similar opinions about the greenhouse effect.

$$df = (R-1)(C-1) = (4-1)(2-1) = 3$$

For α = 0.01 and df = 3, the critical value of χ^2 is 6.251 [From the chi-square distribution table in Appendix-F].

Table E8.6(a)

	Men	Women	Total
Have been big changes	85 (100)	115 (100)	200
Have not been	110 (90)	70 (90)	180
Only small changes	40 (35)	30 (35)	70
Do not know (no opinion)	15 (25)	35 (25)	50
Total	250	250	500

The numbers in parentheses are expected frequencies, which are calculated using the formula:

$$E = \frac{\text{(row total) (column total)}}{\text{(total of both samples)}}$$

The value of the test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 2.25 + 2.25 + 4.4444 + 4.4444 + 0.7143 + 0.7143 + 4 + 4$$

$$= 22.8175$$

Since 22.8175 > 6.251, reject the null hypothesis. Men and women do not have similar opinions about the greenhouse effect.

Example E8.7

A random sample of 100 people selected were asked whether or not they support a certain form of subsidy program from each of the four regions of the United States. Table E8.7 shows the summary of the results obtained. Using the 1% significance level, test the null hypothesis that the percentages of people with different opinions are similar for all four regions.

Table E8.7

	Favour	Oppose	Uncertain
Northeast	55	32	13
Midwest	75	21	4
South	65	28	7
West	60	34	6

SOLUTION:

H_0 : The percentage of people with different opinions are similar for all four regions.

H_1 : The percentage of people with different opinions are *not* similar for all four regions.

$$df = (R-1)(C-1) = (4-1)(3-1) = 6$$

From the chi-square distribution table in Appendix-*F*, the critical value of χ^2 for $df = 6$ and 0.01 area in the right tail is 16.812.

Table E8.7(a)

	Favour	Oppose	Uncertain	Total
Northeast	55 (63.75)	32 (28.75)	13 (7.5)	100
Midwest	75 (63.75)	21 (28.75)	4 (7.5)	100
South	65 (63.75)	28 (28.75)	7 (7.5)	100
West	60 (63.75)	34 (28.75)	6 (7.5)	100
Total	255	115	30	400

The numbers in parentheses are expected frequencies, which are calculated using the formula:

$$E = \frac{\text{(row total) (column total)}}{\text{(total of both samples)}}$$

The value of the test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 1.201 + 0.3674 + 4.0333 + 1.9853 + 2.0891 + 1.6333 + 0.0245 + 0.0196$$

$$+ 0.333 + 0.2206 + 0.9587 + 0.3 = 12.8662$$

Do not reject the null hypothesis. The percentages of people with different opinions are similar for all four regions.

8.5 INFERENCES ABOUT THE POPULATION VARIANCE

The sample variance is a random variable and it possesses a sampling distribution. When all the possible samples of a given size are taken from a population and their variances are calculated, the probability distribution of these variances is called the *sampling distribution of the sample variance*.

If the population from which the sample is selected is approximately normally distributed, then $(n-1)s^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom.

Hence, the chi-square distribution can be used to construct a confidence interval and test a hypothesis about the population variance σ^2.

8.6 ESTIMATION OF THE POPULATION VARIANCE

Considering that the population from which the sample is selected is approximately normally distributed, the $(1-\alpha)100\%$ *confidence interval for the population variance* s^2 is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \text{ to } \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} \qquad (8.7)$$

where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are obtained from the chi-square distribution table in Appendix-*F* for $\alpha/2$ and $1 - \alpha/2$ are as in the right tail of the chi-square distribution curve, respectively, and for $n - 1$ degrees of freedom. The confidence interval for the population standard deviation can be obtained by simply taking the positive square root of the two limits of the confidence interval for the population variance.

Example E8.8 and E8.9 illustrate the estimation of the population variance and population standard deviation.

Example E8.8

Particular machine parts manufactured by a company must have a variance of 0.0025 square millimeters or less. A random sample of 29 such machine parts gave a variance of 0.005 square millimeters.

(*a*) test at the 1% significance level if the variance of all such machine parts is greater than 0.0025 sq mm. Assume that all the machine parts manufactured by the company are approximately normally distributed.

(*b*) make the 98% confidence intervals for the population variance and standard deviation.

SOLUTION:

(*a*) $H_0 : \sigma^2 \leq 0.0025$

$H_1 : \sigma^2 \geq 0.0025$

Area in the right tail = $\alpha = 0.01$

$$df = n - 1 = 29 - 1 = 28$$

χ^2 for $df = 28$ and 0.01 area in the right tail is 48.278 [From the chi-square distribution table in Appendix-*F*].

The value of the test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(29-1)(0.005)}{0.0025} = 56$$

Reject the null hypothesis. The population variance is greater than 0.0025 sq mm.

(*b*) $\frac{\alpha}{2} = \frac{0.02}{2} = 0.01$ and $1 - \frac{\alpha}{2} = 1 - 0.01 = 0.99$

$$df = n - 1 = 29 - 1 = 28$$

χ^2 for 28 *df* and 0.01 area in the right tail = 48.278

χ^2 for 28 *df* and 0.99 area in the right tail = 13.565

The 98% confidence interval for the population variance σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \text{ to } \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} = \frac{(29-1)(0.005)}{48.278} \text{ to } \frac{(29-1)(0.005)}{13.565} = 0.00289987 \text{ to } 0.0103206$$

The 98% confidence interval for σ is

$\sqrt{0.00289987}$ to $\sqrt{0.0103206} = 0.05385$ to 0.10159

Example E8.9

The manufacturer of a certain brand of light bulbs claims that the variance of the lives of these bulbs is 4500 square hours. An independent agency took a random sample of 25 such bulbs and tested them. The variance of the lives of these bulbs was found to be 5000 square hours. Assume that the lives of all such bulbs are approximately normally distributed.

(*a*) make a 99% confidence intervals for the variance and standard deviation of the lives of all such bulbs.

SOLUTION:

(*a*) $df = n - 1 = 25 - 1 = 24$

$\alpha/2 = 0.5 - (0.99/2) = 0.005$ and $1 - \alpha/2 = 1 - 0.005 = 0.995$

From the chi-square distribution table in Appendix-*F*, we have

χ^2 for 24 *df* and 0.005 area in the right tail = 45.559

χ^2 for 24 *df* and 0.995 area in the right tail = 9.886

The 99% confidence interval for the population variance σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \text{ to } \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} = \frac{(25-1)(5000)}{45.559} \text{ to } \frac{(25-1)(5000)}{9.886} = 2633.9472 \text{ to } 12138.3775$$

The 99% confidence interval for σ is

$$\sqrt{2633.9472} \text{ to } \sqrt{12138.3775} = 51.3220 \text{ to } 110.1743$$

8.7 HYPOTHESIS TESTS ABOUT THE POPULATION VARIANCE

The value of the *test statistic* χ^2 is calculated as

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} \tag{8.8}$$

where s^2 = the sample variance

σ^2 = the hypothesised value of the population variance

$n - 1$ = the degrees of freedom

The population from which the sample is selected is assumed to be approximately normally distributed. A test of hypothesis about the population variance can be one-tailed or two-tailed.

Examples E8.10 and E8.11 illustrate the procedure to make tests of hypotheses about σ^2.

Example E8.10

The variance of the SAT scores for all students who took the test this year is 5100. The variance of the SAT scores for a random sample of 20 students from one school is equal to 3220. Test at the 2.5% significance level if the variance of the SAT scores for students from this school is lower than 5100. Assume that the SAT scores for all students at this school are approximately normally distributed.

SOLUTION:

$H_0 : \sigma^2 = 5100$

$H_1 : \sigma^2 < 5100$

Area in the left tail = $\alpha = 0.025$

Area in the right tail = $1 - \alpha = 1 - 0.025 = 0.975$

$df = n - 1 = 20 - 1 = 19$

χ^2 for $df = 19$ and 0.975 area in the right tail = 8.907

[From the chi-square distribution table in Appendix-*F*].

The value of the test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(20-1)(3220)}{5100} = 11.996$$

Do not reject the null hypothesis. The population variance of the SAT scores for students from the given school is not lower than 5100.

Example E8.11

A random sample of 25 students taken from a particular university gave the variance of their GPAs equal to 0.20.

(a) construct the 99% confidence intervals for the population variance and standard deviation. Assume that the GPAs of all students are approximately normally distributed.

(b) the variance of GPAs of all students at this university was 0.12 two years ago. Test at the 1% significance level if the variance of GPAs now is different from 0.12.

SOLUTION:

(a) $\alpha/2 = 0.5 - (0.99/2) = 0.005$ and $1 - \alpha/2 = 1 - 0.005 = 0.995$

$df = n - 1 = 25 - 1 = 24$

From the chi-square distribution table in Appendix-*F*, we have

χ^2 for 24 *df* and 0.005 area in the right tail = 45.559 and χ^2 for 24 *df* and 0.995 area in the right tail = 9.886.

The 99% confidence interval for the population variance σ^2 is

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \text{ to } \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}} = \frac{(25-1)(0.2)}{45.559} \text{ to } \frac{(25-1)(0.20)}{9.886} = 0.10536 \text{ to } 0.48554$$

The 99% confidence interval for σ is

$$\sqrt{0.10536} \text{ to } \sqrt{0.48554} = 0.3246 \text{ to } 0.6968$$

(b) $H_0 : \sigma^2 = 0.12$

$H_1 : \sigma^2 \neq 0.12$

$\alpha/2 = 0.01/2 = 0.005$ and $1 - \alpha/2 = 1 - 0.005 = 0.995$

$df = n - 1 = 25 - 1 = 24$

From the chi-square distribution table in Appendix-*F*, we have

χ^2 for 24 *df* and 0.005 area in the right tail = 45.559 and χ^2 for 24 *df* and 0.995 area in the right tail = 9.886.

The value of the test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(25-1)(0.2)}{0.12} = 40$$

Since 40 is between 9.886 and 45.559, do not reject H_0. The variance of the GPAs is still equal to 0.12.

PROBLEMS

A Goodness-to-Fit Test

P8.1 A randomly selected sample of 100 people who suffer from allergies were asked during what season they suffer the most. Table P8.1 gives the results of this survey.

Table P8.1

Season	Fall	Winter	Spring	Summer
People allergic	19	11	33	37

Using 1% significance level, test the null hypothesis that the proportions of all allergic people are equally distributed over the four seasons.

P8.2 A roulette wheel contains 18 red numbers, 18 black numbers and 2 green numbers. Table P8.2 shows the frequency with which the ball landed on each colour in 200 trials.

Table P8.2

Number	Red	Black	Green
Frequency	88	102	10

At the 5% significance level, do the data suggest that the wheel is out of balance?

P8.3 Table P8.3 lists the age distribution for a sample of 100 people arrested for drunk driving in the year 2007 in a small city in US. Using the 1% significance level, can we reject the null hypothesis that the proportion of people arrested for drunk driving is the same for all age groups?

Table P8.3

Age	18–25	26–35	36–45	46–55	56 and older
Arrests	31	26	20	15	8

P8.4 Table P8.4 shows the results of a survey in 2008 of married woman aged 45 and over when they were asked who makes the financial decisions in their households. In a recent sample of 800 married woman aged 45 and over, 136 indicated that in their households they make the financial decisions themselves, 248 said that their husbands make such decisions, and 416 stated that such decisions are made by both together. Test at the 2.5% significance level whether the current pattern of financial decision-making is different from that for 2008.

Table P8.4

Decision made by	Wife	Husband	Both
Percentage of households	15	36	49

P8.5 Table P8.5 shows the data obtained when the die is rolled 150 times in order to test whether the die is loaded or not (that is, to check whether the six numbers are equally likely or not).

Table P8.5

Number	1	2	3	4	5	6
Frequency	23	26	23	21	31	26

Test at the 0.05 level of significance, do the data provide sufficient evidence to conclude that the die is loaded?

Contingency Tables

P8.6 A random sample of 300 adults was selected and they were asked if they favour giving more freedom to elementary school teachers to approximately punish students for lack of discipline. The results are summarised in Table P8.6.

Table P8.6

	In favour (F)	Against (A)	No opinion (N)
Men (M)	95	68	12
Women (W)	85	34	6

Find the expected frequencies for this Table P8.6 assuming that the two attributes, gender and opinions are independent on the issue.

P8.7 Refer to Problem P8.6. Use a 1% significance level to determine whether the sample provide sufficient information to conclude that the two attributes, gender and opinion of adults, are dependent.

P8.8 The results obtained as shown in Table P8.8 from a survey of college students who were interviewed on their drinking behaviour. Students were classified according to their academic performance at the college. The students were asked the question whether or not he/she had engaged in *binge drinking* (five or more drinks in a row) at least once during the past month.

Table P8.8

College performance	Binge drinking	
	Yes	No
Excellent	1265	3600
Average	2200	4200
Below average	450	500

Using the 5% significance level, can we conclude that college performance and binge drinking are related for all college students?

P8.9 Table P8.9 gives a two-way classification of divorced fathers based on a legal study and compliance with child support court orders. At the 1% significance level, can we conclude that the legal custody and compliance with child support orders are related?

Table P8.9

Legal custody	Number of fathers who		
	Pay nothing	Pay in past	Pay in full
Mother	400	1100	1000
Joint/split	75	230	400

P8.10 Two random samples, one of 100 administrators and a second of 50 engineers were taken from a large company. These employees were asked about their views on a certain company issue. Table P8.10 gives the results of this survey.

Table P8.10

	Opinion		
	Favour	Oppose	Uncertain
Administrators	61	30	9
Engineers	20	247	3

Use the 2.8% significance level, test the null hypothesis that the distribution of opinions are homogeneous for the two groups of workers.

A Test of Independence or Homogeneity

P8.11 Table P8.11 gives the data on income distribution for households in Visakhapatnam and Hyderabad. Using the 2.5% significance level, test the null hypothesis that the distribution of households with regard to income levels is similar (homogeneous) for the two cities.

Table P8.11

	Visakhapatnam	Hyderabad	Total
High income	70	34	104
Medium income	81	39	120
Low income	99	77	176
Total	250	150	400

P8.12 Table P8.12 shows from a sample of 500 people who held more than one job to make ends meet.

Table P8.12

	Single	Married	Other
Male	70	211	39
Female	33	101	46

Test at the 10% significance level if gender and marital status are related for all people who hold more than one job.

P8.13 A random sample of 100 people was selected from each of the four regions of the state of Andhra Pradesh. These people were asked whether or not they support a certain farm subsidy programme. The results of the survey are summarised in Table P8.13.

Table P8.13

	Favour	Oppose	No opinion
Northeast	55	34	11
Midwest	73	22	5
South	68	27	5
West	59	34	7

Using the 1% significance level, test the null hypothesis that the percentages of people with different opinions are similar for all four groups.

P8.14 Two drugs were given to two randomly selected patients to cure the same disease. Table P8.14 gives the data about the number of patients who were cured and not cured by each of the two drugs.

Table P8.14

	Cured	Not cured
Drug I	47	13
Drug II	16	24

Test at the 1% significance level whether or not the two drugs are similar in curing and not curing the patients.

P8.15 Table P8.15 gives the results of a survey of two samples taken from a company workers. The workers were asked about their views on health care programme. Using the 2.5% significance level, test the null hypothesis that the distributions of opinions are homogeneous for the two groups of workers.

Table P8.15

	Opinion		
	Favour	Oppose	No opinion
Male workers	46	40	9
Female workers	20	27	3

Inferences about the Population Variance

P8.16 One type of machine component is manufactured by a company. The machine that fills packages of these parts is set up in a way that the average net weight of these packages is 32 gms with a variance of 0.015 sq.gms. Periodically, the quality control inspector selects a sample of a few such packages to construct a 95% confidence interval for the population variance. If either of both or one of the two limits of the confidence interval is not in the interval 0.008 to 0.030, the machine is stopped and adjusted. A random sample of 25 packages from the population gave a sample variance of 0.030 sq.gms. Based on this sample data, do we say that the machine needs an adjustment? Assume that the net weights of all parts in all packages are normally distributed.

P8.17 A sample of 16 observations selected from a normally distributed population produced a sample variance of 1.11. Write the null and alternative hypothesis to test whether the population variance is greater than 0.75. For $\alpha = 0.01$, find the critical value of χ^2. Using the 1% significance level, can we reject the null hypothesis?

P8.18 A sample of 14 observations selected from a normally distributed population produced a sample variance of 4.7. Write the null and alternative hypothesis to test whether the population variance is different from 2.3. Using $\alpha = 0.05$, find the critical values of χ^2. Using the 5% significance level, can we reject the null hypothesis stated in above?

P8.19 Steel bolts for a machine manufactured by a company must have a variance of 0.0025 sq.mm or less for acceptance y a customer. A random sample of 29 such bolts gave a variance of 0.0058 sq.mm. Assume that the lengths of all bolts manufactured by this company are approximately normally distributed. Test at the 1% significance level if the variance of all such bolts is greater than 0.0025 sq.mm.

P8.20 The manufacturer of a certain brand of light bulbs claims that the variance of the lives of these bulbs is 4500 square hours. In a random sample of 25 such bulbs, the variance of the lives of these bulbs was found to be 4800 square hours. Assume that the lives of all such bulbs are approximately normally distributed. Test at the 5% significance level if the variance of such light bulbs is different from 4500 square hours.

Estimation of the Population Variance

P8.21 Refer to Problem P8.19. Make a 98% confidence intervals for the population variance and standard deviation.

P8.22 Refer to Problem P8.20. Make the 99% confidence intervals for the variance and standard deviation of the lives of all such bulbs.

P8.23 Refer to Problem P8.16. Use $\alpha = 0.01$. Do we think that the machine needs an adjustment?

P8.24 The variance scores on a standardised science test for all high school seniors was 150 in the year 2008. A sample score for 20 high school seniors who took this test in the year 2008 gave a variance of 173. Test at the 5% significance level if the variance of current scores of all high school seniors on this test is different from 150. Assume that the scores of all high school seniors on this test are approximately normally distributed.

REVIEW QUESTIONS

1. Why is the phrase "goodness-of-fit" used to describe the type of hypothesis test considered in this chapter?
2. Are the observed frequencies variables? Explain your answers.
3. Are the expected frequencies variables? Explain your answers.
4. To find out whether two variables of a population are associated, we usually need to resort to inferential methods such as chi-square independence test. Why?
5. Explain why a chi-square goodness-of-fit test or a chi-square independence test is always right tailed.

6. If the observed and expected frequencies for a chi-square goodness-of-fit test or a chi-square independence test matched perfectly, what would be the value of the test statistic?
7. State the assumptions for a chi-square goodness-of-fit test or a chi-square independence test regarding the expected frequency.
8. Describe the four characteristics of a multinomial experiment.
9. When is a goodness-of-fit test applied? Explain.
10. What is the difference between the observed and expected frequencies for a goodness-of-fit test?
11. How is the expected frequency of a category computed for a goodness-of-fit test?
12. Describe the test of independence.
13. Describe the test of homogeneity.
14. How are the expected frequencies for cells of a contingency table calculated in a test of independence or homogeneity?

STATE TRUE OR FALSE

1. The random variable χ^2 assumes only non-negative values. (True/False)
2. The parameters of the chi-square distribution are *df* and *n*. (True/False)
3. The trials are random is a characteristic of a multinomial experiment. (True/False)
4. The observed frequencies for a goodness-of-fit test are the frequencies obtained from the performance of an experiment. (True/False)
5. The expected frequencies for a goodness-of-fit test are the frequencies given by the product of *n* and *p*. (True/False)
6. The degrees of freedom for a goodness-of-fit test are $n + k - 1$. (True/False)
7. The chi-square goodness-of-fit test is always right tailed. (True/False)
8. The expected frequency of each category must be at least 8 to apply a goodness-of-fit test. (True/False)
9. The degrees of freedom for a test of independence are $(R - 1)(C - 1)$. (True/False)
10. The trials are independent in a multinomial experiment. (True/False)
11. The probabilities of the various outcomes remain constant for each trial in a multinomial experiment. (True/False)
12. The value of the test statistic χ^2 for a test of independence is calculated as $\chi^2 = \sum \frac{(O-E)^2}{E}$ (True/False)
13. The test of independence is a test of null hypothesis that two attributes of a population are related. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **False** 3. **False** 4. **True** 5. **True** 6. **False** 7. **True** 8. **False** 9. **True** 10. **True**
11. **True** 12. **True** 13. **False**

CHAPTER 9

Analysis of Variance

9.1 INTRODUCTION

In Chapter 5, we studied how to compare two population means, that is, the means of a single variable for two different populations. We reviewed various methods for making such comparisons. We described the procedures that are used to test hypotheses about the difference between two population means using the normal and t-distributions. The hypothesis testing procedures for the difference between two population proportions using the normal distribution were also described. In Chapter 6, the procedures to test hypotheses about the equally of more than two population proportions using the chi-square distribution were also presented. Analysis of variance (ANOVA) provides methods for comparing several population means, that is, the means of a single variable for several populations. In this chapter, we present the simplest kind of ANOVA, *one-way analysis of variance*. This type of ANOVA is called *one-way* analysis of variance because it compares the means of a variable for populations that result from a classification by *one* other variable, called the *factor*. The possible values of the factor are referred to as the *levels* of the factor. One-way analysis of variance is the generalisation to more than two populations of pooled t-procedure.

An *analysis of variance* test is performed using the F-distribution. This chapter explains how to test the null hypothesis that the means of more than two populations are equal.

9.2 ONE-WAY ANALYSIS OF VARIANCE

ANOVA is a procedure used to test the null hypothesis that the means of three or more populations are equal. In this section, we discuss the *one-way* ANOVA procedure to make to make tests comparing the means of several populations. By using one-way ANOVA test, we analyse only one factor or variable. The procedure under discussion is called the analysis of variance because the test is based on the analysis of variation in the data obtained from different samples. The application of one-way ANOVA requires the following assumptions.

Assumptions (conditions) for one-way ANOVA

1. **Simple random samples:** The samples taken from the populations under consideration are simple random samples.

2. **Independent samples:** The samples taken from the populations under consideration are independent of one another.
3. **Normal populations:** For each population, the variable under consideration is normally distributed.
4. **Equal standard deviations:** The standard deviations of the variable under consideration are the same for all the populations.

The reason for the word *variance* in *analysis of variance* is that the procedure for comparing the means analyses the variation in the sample data.

The one-way ANOVA test is always *right-tailed with the rejection region in the right-tail of the F-distribution curve.*

9.3 TEST STATISTIC

The value of the test statistic F for a test of hypothesis using ANOVA is given by the ratio of two variances, the variance between samples (MSB) and the variance within samples (MSW). The value of the test statistic F for an ANOVA test is given by

$$F = \frac{\text{variance between samples}}{\text{variance within samples}} \text{ or } \frac{\text{MSB}}{\text{MSW}} \tag{9.1}$$

The *between-samples sum of squares*, denoted by SSB is calculated as

$$\text{SSB} = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \dots + \frac{T_n^2}{n_n}\right) - \frac{(\Sigma x)^2}{n} \tag{9.2}$$

The within samples sum of squares, denoted by SSW, is calculated as

$$\text{SSW} = \Sigma x^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \dots + \frac{T_n^2}{n_n}\right) \tag{9.3}$$

where n_i = the size of the sample

T_i = the sum of the values in sample i

n = the number of values in all samples = $n_1 + n_2 + \cdots$

Σx = the sum of the values in all samples = $T_1 + T_2 + T_3 + \cdots$

Σx^2 = the sum of the squares of the values in all samples.

The total sum of squares, denoted by SST, is calculated as

$$\text{SST} = \text{SSB} + \text{SSW} \tag{9.4}$$

The MSB and MSW are calculated as $MSB = \dfrac{SSB}{k-1}$

and $$\text{MSW} = \frac{\text{SSW}}{n-k} \tag{9.5}$$

where $k - 1$ and $n - k$ are, respectively, the degrees of freedom for the numerator and the degrees of freedom for the denominator for the F-distribution.

Example E9.1

Table E9.1 gives the number of classes missed during one semester by 25 randomly selected undergraduate continuing education students drawn from three different groups.

Table E9.1

Below 25	25 to 30	31 and above
18	10	5
11	5	7
26	11	5
11	15	4
19	7	9
4	9	9
16	4	
12	12	
17	18	
9		

(*a*) state the null and alternative hypotheses if we were to test if the mean number of classes missed during the semester by all students in each of these three age groups is the same

(*b*) show the rejection and non-rejection regions on the *F*-distribution curve for $\alpha = 0.01$

(*c*) find SSB, SSW and SST

(*d*) find the between-samples and within-samples variances

(*e*) find the calculated value of the test statistic *F*

(*f*) show the ANOVA table for this problem.

SOLUTION:

(*a*) $H_0: \mu_1 = \mu_2 = \mu_3$

H_1: All three population means are not equal

(*b*) $k = 3$ and $n = n_1 + n_2 + n_3 = 10 + 9 + 6 = 25$

df for the numerator $= k - 1 = 3 - 1 = 2$

df for the denominator $= n - k = 25 - 3 = 22$

For $\alpha = 0.01$ and $df = (2, 22)$, the critical value of *F* is 5.72.

The rejection region lies to the right of $F = 5.72$. The non-rejection region lies to the left of $F = 5.72$ [From the *F*-distribution table in Appendix-*H*].

(*c*) $T_1 = 143$, $T_2 = 91$ and $T_3 = 39$

$\Sigma x = T_1 + T_2 + T_3 = 143 + 91 + 39 = 273$

$\Sigma x^2 = 18^2 + 11^2 + 26^2 + 11^2 + 19^2 + 4^2 + 16^2 + 12^2 + 17^2 + 9^2 + 10^2 + 5^2 + 11^2 + 15^2 + 7^2 + 9^2 + 4^2 + 12^2 + 18^2 + 5^2 + 7^2 + 5^2 + 9^2 + 9^2 = 3751$

$$SSB = \left[\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3}\right] - \frac{(\Sigma x)^2}{n} = \left[\frac{143^2}{10} + \frac{91^2}{9} + \frac{39^2}{6}\right] - \frac{273^2}{25} = 237.3511$$

$$\text{SSW} = \Sigma x^2 - \left[\frac{T_1^2}{n_1}+\frac{T_2^2}{n_2}+\frac{T_3^2}{n_3}\right] = 3751 - \left[\frac{143^2}{10}+\frac{91^2}{9}+\frac{39^2}{6}\right] = 532.4889$$

$$\text{SST} = \text{SSB} + \text{SSW} = 237.3511 + 532.4889 = 769.84$$

(*d*) df for the numerator $= k - 1 = 3 - 1 = 2$

df for the denominator $= n - k = 25 - 3 = 22$

$\text{MSB} = \text{SSB}/(k-1) = 237.3511/(3-1) = 118.6756$

$\text{MSW} = \text{SSW}/(n-k) = 532.4889/(25-3) = 24.2040$

(*e*) For $\alpha = 0.01$ and $df = (2, 22)$, the critical value of $F = 5.72$ as shown in Fig. E9.1.

The value of the test statistic is

$F = \text{MSB/MSW} = 24.2040/118.6756 = 4.9031$

(*f*)

ANOVA Table E9.1(a)

Source of variation	Degrees of freedom	Sum of squares	Mean square	Value of the test statistic
Between	2	237.3511	118.6756	$F = \dfrac{118.6756}{24.2040} = 4.9031$
Within	22	532.4889	24.2040	
Total	24	769.84		

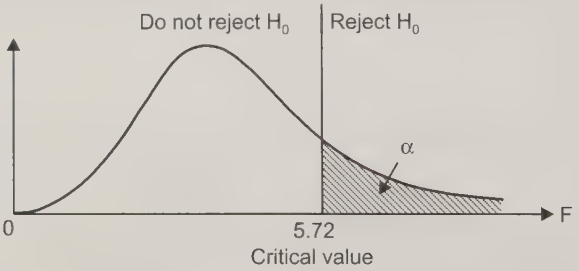


Fig. E9.1

9.4 ONE-WAY ANOVA TEST

Here, we test the null hypothesis that the means of all groups are all equal for all the population samples against the alternative hypothesis that the means of all the groups are not equal. In a one-way ANOVA test, the null hypothesis is that the means for all populations are equal. The alternative hypothesis states that at least one of the population means is different from the others. Table 9.1 gives a step-by-step method for conducting a one-way ANOVA test by using the critical value approach. Because the null hypothesis is rejected only when the test statistic, *F*, is too large, a one-way ANOVA test is always right tailed.

Table 9.1: One-way ANOVA test

Purpose: To perform a hypothesis test to compare k population means, $\mu_1, \mu_2, \ldots, \mu_n$

Assumptions:

1. Simple random samples
2. Independent samples
3. Normal populations
4. Equal population standard deviations

Step 1: The null and alternative hypotheses are

H_0: $\mu_1 = \mu_2 = \ldots = \mu_n$

H_1: Not all the means are equal

Step 2: Decide on the significance level, α

Step 3: Obtain the three sums of squares, SSB, SSW and SST

Step 4: Construct a one-way ANOVA table to find the value of the F-statistic

Source of variation	*df*	Sum of squares	Mean squares = SS/*df*	F-statistic (Test statistic)
Between Within	$k - 1$ $n - k$	SSB SSW	MSB MSW	$F = \dfrac{\text{MSB}}{\text{MSW}}$
Total	$n - 1$	SST		

Step 5: The critical value is F_c with $df = (k - 1, n - k)$

Step 6: If the value of the F-statistic falls in the region of rejection, reject H_0; otherwise, do not reject H_0.

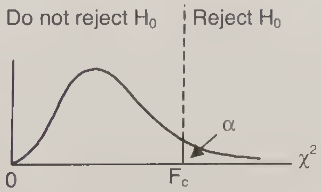


Fig. 9.1

Step 7: Interpret the results of the hypothesis test.

Example E9.2

Refer to Example E9.1. Can we reject the null hypothesis stated in Example E9.1 in part *(a)* at a significance level of 1%?

SOLUTION:

Refer to the solution given in Example E9.1.

Since $F = 4.9031 < 5.72$, do not reject H_0. The mean number of classes missed by all three age groups is the same.

Example E9.3

Table E9.3 shows the life of the light bulbs (in thousands) of four different brands investigated by a company in order to find out if the mean life of the four brands of light bulbs is the same.

Table E9.3

Brand A	Brand B	Brand C	Brand D
22	20	23	25
23	22	26	24
20	19	25	22
25	23	27	29
22	20	23	28
23	21	19	25
27	20	29	27

Test the null hypothesis at the 2.5% significance level that the mean life of light bulbs for each of these four brands is the same.

SOLUTION:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: The means of the four populations are not equal

$k = 4$ and $n = n_1 + n_2 + n_3 + n_4 = 7 + 7 + 7 + 7 = 28$

df for the numerator $= k - 1 = 4 - 1 = 3$

df for the denominator $= n - k = 28 - 4 = 24$

For $\alpha = 0.025$ and $df = (3, 24)$, the critical value of F is 3.72 [From the F-distribution table in Appendix-F].

$T_1 = 162, T_2 = 145, T_3 = 172, T_4 = 180$

$\Sigma x = T_1 + T_2 + T_3 + T_4 = 162 + 142 + 172 + 180 = 659$

$\Sigma x^2 = (22^2 + 23^2 + \cdots + 27^2 + 20^2 + 22^2 + \cdots + 20^2 + 23^2 + 26^2 + \cdots + 29^2 + 25^2 + 24^2 + \cdots + 27^2)$
$= 15{,}749$

$$\text{SSB} = \left[\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4}\right] - \frac{(\Sigma x)^2}{n} = 97.5357$$

$$\text{SSW} = \Sigma x^2 - \left[\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4}\right] = 141.4286$$

$\text{MSB} = \text{SSB}/\,(k - 1) = 97.5357/\,(4 - 1) = 32.5119$

$\text{MSW} = \text{SSW}/\,(n - k) = 141.4286/\,(28 - 4) = 5.8929$

The value of the test statistic is

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{32.5119}{5.8929} = 5.5172$$

Reject H_0. The mean life of light bulbs for each of these four brands A, B, C and D is not the same.

PROBLEMS

P9.1 Table P9.1 shows the test scores of the students in three groups of fifteen students selected at random and who were experimented with three different methods of teaching algebra. Compute the value of the test statistic *F*.

Table P9.1

Method A	Method B	Method C
47	60	85
72	83	67
51	68	94
64	70	74
88	89	68

P9.2 Refer to Problem P9.1. At the 1% significance level, can we reject the null hypothesis that the mean algebra score of all students taught by each of these three methods *A*, *B* and *C* is the same?

P9.3 A study conducted gave the number of customers served by four bank tellers during each of the observed hours is summarised in Table P9.3. At the 5% significance level, test the null hypothesis that the mean number of customers served per hour by each of these four tellers is the same. Use the one-way ANOVA procedure.

Table P9.3

Teller A	Teller B	Teller C	Teller D
20	15	11	23
19	16	15	20
25	13	20	21
23	12	13	25
19	17	15	19
	13	19	

P9.4 Table P9.4 shows the data obtained for two samples selected at random from two populations that are independent and normally distributed with equal variances.

Table P9.4

Sample A	Sample B
15	10
16	9
12	12
9	17
12	15
19	8
17	9

Using the one-way ANOVA procedure, test at the 5% significance level whether the means of the populations from which these samples are drawn are equal.

P9.5 Table P9.5 gives the production of wheat (in bushels) for each acre for three brands of fertilizer. At the 5% significance level, can we conclude that the mean yield of wheat for each of these three brands of fertilizer is the same?

Table P9.5

Brand A	Brand B	Brand C
71	57	61
70	43	57
74	54	62
57	48	71
63	46	55
67	53	66
71	48	57
68	57	64

P9.6 Table P9.6 lists the weight (in kg) gained by fifteen chickens randomly put in each of the three groups and fed by three brands of diet. At the 1% significance level, can we conclude that the mean weight gained by all chickens is the same for each of these three diets?

Table P9.6

Brand A	Brand B	Brand C
0.4	0.3	0.6
0.6	0.65	0.4
0.8	0.31	0.34
0.5	0.23	0.74
0.4	0.35	0.44

P9.7 A consumer agency that wanted to compare the drying times for paints made of three brands tested a few samples of paints from each of these three brands. Table P9.7 records the drying times (in minutes) for these samples of paints.

Table P9.7

Brand A	Brand B	Brand C
41	56	45
52	62	46
39	61	52
48	55	59
41	48	43
52	61	41
55		46

Using the 5% significance level, test the null hypothesis that the mean drying times for paints of these three brands are equal.

P9.8 A consumer agency wants to check if the mean lives of four brands of light bulbs, which sell for nearly the same price, are the same. The agency randomly selected a few light bulbs of each brand and tested them. Table P9.8 gives the lives of these light bulbs in thousands of hours.

Table P9.8

Brand A	Brand B	Brand C	Brand D
73	52	58	57
67	54	72	52
51	47	80	43
57	56	69	44
65		65	

At the 5% significance level, can we reject the null hypothesis that the mean life of each of these four brands of listed bulbs is the same?

P9.9 Table P9.9 gives the independent random samples of households in the four US regions giving the last year's energy consumptions. At the 5% significance level, does the data provide sufficient evidence to conclude that a difference exists in the last year's mean energy consumption by households among the four US regions?

Table P9.9

Northeast	Midwest	South	West
14	16	12	9
11	13	8	11
12	17	9	8
14	13	13	7
13	14		9
	12		

P9.10 Table P9.10 shows the data from independent simple random samples from several populations. Decide at the 5% significance level, whether the data provide sufficient evidence to conclude that the means of the populations from which the samples were drawn are not all the same.

Table P9.10

Sample 1	Sample 2	Sample 3	Sample 4
6	9	4	8
3	6	4	5
4	7	3	6
	5	2	
	6	3	

REVIEW QUESTIONS

1. Describe the basic properties of F-curves in the F-distribution.
2. State the assumptions (conditions) made in one-way ANOVA.
3. Explain when a one-way ANOVA procedure is used to make a test of hypothesis.
4. Explain the reason for the word *variance* in the phrase *analysis of variance*.
5. In a one-way ANOVA, identify the statistic used
   - *(a)* as a measure of variance among the sample means
   - *(b)* as a measure of variance within the samples
   - *(c)* to compare the variance among the sample means to the variation within the samples.

6. Describe the logic behind one-way ANOVA.
7. What does the term *one-way* signify in the phrase *one-way* ANOVA?
8. In a one-way ANOVA, what is the residual of an observation?
9. Following are the notations for the three sums of squares. State the name of each sum of squares and the source of variation each sum of squares represents.
    (*a*) SSB (*b*) SST (*c*) SSW
10. State the one-way ANOVA identity and interpret its meaning with regard to partitioning the total variation in the data.
11. In each part, specify what type of analysis one might use?
    - (*a*) To study the effect of one factor on the mean of a response variable.
    - (*b*) To study the effect of two factors on the mean of a response variable.
    - (*c*) If we define $s = \sqrt{\text{MSW}}$, of which parameter is s an estimate?
12. True or false: If we know any two of the three sums of squares, SST, SSB and SSW, can we determine the remaining one? Explain your answer.

STATE TRUE OR FALSE

1. In one-way ANOVA, the populations from which the samples are drawn are approximately normally distributed. (True/False)
2. The population from which the samples are drawn have the same variance (or standard deviation). (True/False)
3. The samples drawn from different populations are random and dependent. (True/False)
4. The variance between samples is also called the mean square between samples. (True/False)
5. The variance within samples is also called the mean square within samples. (True/False)
6. ANOVA is a procedure used to test the null hypothesis that the means of three or more populations are equal. (True/False)
7. The value of the *Test statistic F* for an ANOVA test is calculated as $F = \dfrac{\text{variance between samples}}{\text{variance within samples}}$ (True/False)
8. The MSB is calculated as $\text{MSB} = \dfrac{\text{SSB}}{n-k}$ (True/False)
9. The MSW is calculated as $\text{MSW} = \dfrac{\text{SSW}}{n-k}$ (True/False)
10. The analysis of variance technique that analyses one variable only. (True/False)
11. Mean square between samples is a measure of the variation among means of samples taken from different populations. (True/False)
12. Mean square within samples is a measure of the variation within data of all samples taken from different populations. (True/False)
13. The *F*-distribution is continuous. (True/False)
14. The *F*-distribution is always skewed to the left. (True/False)

15. The units of the F-distribution are always positive. (True/False)
16. The one-way ANOVA test analyses only one population. (True/False)
17. The one-way ANOVA test is always right-tailed. (True/False)
18. For a one-way ANOVA with k treatments and n observations in all samples taken together, the number of degrees of freedom for the numerator are $k - 1$. (True/False)
19. For a one-way ANOVA with k treatments and n observations in all samples taken together, the number of degrees of freedom for the denominator are $n - k$. (True/False)
20. The ANOVA test can be applied to compare more than four population means only. (True/False)

ANSWERS TO STATE TRUE OR FALSE

1. **True** 2. **True** 3. **False** 4. **True** 5. **True** 6. **True** 7. **True** 8. **False** 9. **True** 10. **True**
11. **True** 12. **True** 13. **True** 14. **False** 15. **False** 16. **False** 17. **True** 18. **True** 19. **True** 20. **False**

❐ ❐ ❐

Appendices

Appendix–A

Values for *n* Factorial

n	$n!$	n	$n!$	n	$n!$
0	1 (by definition)	40	8.15915×10^{47}	80	7.15695×10^{118}
1	1	41	3.34525×10^{49}	81	5.79713×10^{120}
2	2	42	1.40501×10^{51}	82	4.75364×10^{122}
3	6	43	6.04153×10^{52}	83	3.94552×10^{124}
4	24	44	2.65827×10^{54}	84	3.31424×10^{126}
5	120	45	1.19622×10^{56}	85	2.81710×10^{128}
6	720	46	5.50262×10^{57}	86	2.42271×10^{130}
7	5040	47	2.58623×10^{59}	87	2.10776×10^{132}
8	40,320	48	1.24139×10^{61}	88	1.85483×10^{134}
9	362,880	49	6.08282×10^{62}	89	1.65080×10^{136}
10	3,628,800	50	3.04141×10^{64}	90	1.48572×10^{138}
11	39,916,800	51	1.55112×10^{65}	91	1.35200×10^{140}
12	479,001,600	52	8.06582×10^{67}	92	1.24384×10^{142}
13	6,227,020,800	53	4.27488×10^{69}	93	1.15677×10^{144}
14	87,178,291,200	54	2.30844×10^{71}	94	1.08737×10^{146}
15	1,307,674,368,000	55	1.26964×10^{73}	95	1.03300×10^{148}
16	20,922,789,888,000	56	7.10999×10^{74}	96	9.91678×10^{149}
17	355,687,428,096,000	57	4.05269×10^{76}	97	9.61928×10^{151}
18	6,402,373,705,728,000	58	2.35056×10^{78}	98	9.42689×10^{153}
19	121,645,100,408,832,000	59	1.38683×10^{80}	99	9.33262×10^{155}
20	2,432,902,008,176,640,000	60	8.32099×10^{81}	100	9.33262×10^{157}
21	51,090,942,171,709,440,000	61	5.07580×10^{83}		
22	1,124,000,727,777,607,680,000	62	3.14700×10^{85}		
23	25,852,016,738,884,976,640,000	63	1.98261×10^{87}		
24	620,448,401,733,239,439,360,000	64	1.26887×10^{89}		
25	15,511,210,043,330,985,984,000,000	65	8.24765×10^{90}		
26	403,291,461,126,605,635,584,000,000	66	5.44345×10^{92}		
27	10,888,869,450,418,352,160,768,000,000	67	3.64711×10^{94}		
28	304,888,344,611,713,860,501,504,000,000	68	2.48004×10^{96}		
29	8,841,761,993,739,701,954,543,616,000,000	69	1.71122×10^{98}		
30	265,252,859,812,191,058,636,308,480,000,000				
31	8.22284×10^{33}	70	1.19786×10^{100}		
32	2.63131×10^{35}	71	8.50479×10^{101}		
33	8.68332×10^{36}	72	6.12345×10^{103}		
34	2.95233×10^{38}	73	4.47012×10^{105}		
35	1.03331×10^{40}	74	3.30789×10^{107}		
36	3.71993×10^{41}	75	2.48091×10^{109}		
37	1.37638×10^{43}	76	1.88549×10^{111}		
38	5.23023×10^{44}	77	1.45183×10^{113}		
39	2.03979×10^{46}	78	1.13243×10^{115}		
		79	8.94618×10^{116}		

Appendix–B

Binomial Coefficients

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!} = \binom{n}{n-k}, 0!=1$$

For $k > 20$, use the fact that $\binom{n}{k} = \binom{n}{n-k}$

n k	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1																			
2	1	2	1																		
3	1	3	3	1																	
4	1	4	6	4	1																
5	1	5	10	10	5	1															
6	1	6	15	20	15	6	1														
7	1	7	21	35	35	21	7	1													
8	1	8	28	56	70	56	28	8	1												
9	1	9	36	84	126	126	84	36	9	1											
10	1	10	45	120	210	252	210	120	45	10	1										
11	1	11	55	165	330	462	462	330	165	55	11	1									
12	1	12	66	220	495	792	924	792	495	220	66	12	1								
13	1	13	78	286	715	1,287	1,716	1,716	1,287	715	286	78	13	1							
14	1	14	91	364	1,001	2,002	3,003	3,432	3,003	2,002	1,001	364	91	14	1						
15	1	15	105	455	1,365	3,003	5,005	6,435	6,435	5,005	3,003	1,365	455	105	15	1					
16	1	16	120	560	1,820	4,368	8,008	11,440	12,870	11,440	8,008	4,368	1,820	560	120	16	1				
17	1	17	136	680	2,380	6,188	12,376	19,448	24,310	24,310	19,448	12,376	6,188	2,380	680	136	17	1			
18	1	18	153	816	3,060	8,568	18,564	31,824	43,758	48,620	43,758	31,824	18,564	8,568	3,060	816	153	18	1		
19	1	19	171	969	3,876	11,628	27,132	50,388	75,582	92,378	92,378	75,582	50,388	27,132	11,628	3,876	969	171	19	1	
20	1	20	190	1,140	4,485	15,504	38,760	77,520	125,970	167,960	184,756	167,960	125,970	77,520	38,760	15,504	4,845	1,140	190	20	1
21	1	21	210	1,330	5,985	20,349	54,264	116,280	203,490	293,930	352,716	352,716	293,930	203,490	116,280	54,264	20,349	5,985	1,330	210	21
22	1	22	231	1,540	7,315	26,334	74,613	170,544	319,770	497,420	646,646	705,432	646,646	497,420	319,770	170,544	74,613	26,334	7,315	1,540	231
23	1	23	253	1,771	8,855	33,649	100,947	245,157	490,314	817,190	1,144,066	1,352,078	1,352,078	1,144,066	817,190	490,314	245,157	100,947	33,649	8,855	1,771
24	1	24	276	2,024	10,626	42,504	134,596	346,104	735,471	1,307,504	1,961,256	2,496,144	2,704,156	2,496,144	1,961,256	1,307,504	735,471	346,104	134,596	42,504	10,626
25	1	25	300	2,300	12,650	53,130	177,100	480,700	1,081,575	2,042,975	3,268,760	4,457,400	5,200,300	5,200,300	4,457,400	3,268,760	2,042,975	1,081,575	480,700	177,100	53,130

Appendix–C

Table of Binomial Probabilities

		p										
n	*x*	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
1	0	.9500	.9000	.8000	.7000	.6000	.5000	.4000	.3000	.2000	.1000	.0500
	1	.0500	.1000	.2000	.3000	.4000	.5000	.6000	.7000	.8000	.9000	.9500
2	0	.9025	.8100	.6400	.4900	.3600	.2500	.1600	.0900	.0400	.0100	.0025
	1	.0950	.1800	.3200	.4200	.4800	.5000	.4800	.4200	.3200	.1800	.0950
	2	.0025	.0100	.0400	.0900	.1600	.2500	.3600	.4900	.6400	.8100	.9025
3	0	.8574	.7290	.5120	.3430	.2160	.1250	.0640	.0270	.0080	.0010	.0001
	1	.1354	.2430	.3840	.4410	.4320	.3750	.2880	.1890	.0960	.0270	.0071
	2	.0071	.0270	.0960	.1890	.2880	.3750	.4320	.4410	.3840	.2430	.1354
	3	.0001	.0010	.0080	.0270	.0640	.1250	.2160	.3430	.5120	.7290	.8574
4	0	.8145	.6561	.4096	.2401	.1296	.0625	.0256	.0081	.0016	.0001	.0000
	1	.1715	.2916	.4096	.4116	.3456	.2500	.1536	.0756	.0256	.0036	.0005
	2	.0135	.0486	.1536	.2646	.3456	.3750	.3456	.2646	.1536	.0486	.0135
	3	.0005	.0036	.0256	.0756	.1536	.2500	.3456	.4116	.4096	.2916	.1715
	4	.0000	.0001	.0016	.0081	.0256	.0625	.1296	.2401	.4096	.6561	.8145
5	0	.7738	.5905	.3277	.1681	.0778	.0312	.0102	.0024	.0003	.0000	.0000
	1	.2036	.3280	.4096	.3602	.2592	.1562	.0768	.0284	.0064	.0005	.0000
	2	.0214	.0729	.2048	.3087	.3456	.3125	.2304	.1323	.0512	.0081	.0011
	3	.0011	.0081	.0512	.1323	.2304	.3125	.3456	.3087	.2048	.0729	.0214
	4	.0000	.0004	.0064	.0283	.0768	.1562	.2592	.3601	.4096	.3281	.2036
	5	.0000	.0000	.0003	.0024	.0102	.0312	.0778	.1681	.3277	.5905	.7738
6	0	.7351	.5314	.2621	.1176	.0467	.0156	.0041	.0007	.0001	.0000	.0000
	1	.2321	.3543	.3932	.3025	.1866	.0937	.0369	.0102	.0015	.0001	.0000
	2	.0305	.0984	.2458	.3241	.3110	.2344	.1382	.0595	.0154	.0012	.0001
	3	.0021	.0146	.0819	.1852	.2765	.3125	.2765	.1852	.0819	.0146	.0021
	4	.0001	.0012	.0154	.0595	.1382	.2344	.3110	.3241	.2458	.0984	.0305
	5	.0000	.0001	.0015	.0102	.0369	.0937	.1866	.3025	.3932	.3543	.2321
	6	.0000	.0000	.0001	.0007	.0041	.0156	.0467	.1176	.2621	.5314	.7351
7	0	.6983	.4783	.2097	.0824	.0280	.0078	.0016	.0002	.0000	.0000	.0000
	1	.2573	.3720	.3670	.2471	.1306	.0547	.0172	.0036	.0004	.0000	.0000
	2	.0406	.1240	.2753	.3177	.2613	.1641	.0774	.0250	.0043	.0002	.0000
	3	.0036	.0230	.1147	.2269	.2903	.2734	.1935	.0972	.0287	.0026	.0002
	4	.0002	.0026	.0287	.0972	.1935	.2734	.2903	.2269	.1147	.0230	.0036
	5	.0000	.0002	.0043	.0250	.0774	.1641	.2613	.3177	.2753	.1240	.0406
	6	.0000	.0000	.0004	.0036	.0172	.0547	.1306	.2471	.3670	.3720	.2573
	7	.0000	.0000	.0000	.0002	.0016	.0078	.0280	.0824	.2097	.4783	.6983
8	0	.6634	.4305	.1678	.0576	.0168	.0039	.0007	.0001	.0000	.0000	.0000
	1	.2793	.3826	.3355	.1977	.0896	.0312	.0079	.0012	.0001	.0000	.0000
	2	.0515	.1488	.2936	.2965	.2090	.1094	.0413	.0100	.0011	.0000	.0000
	3	.0054	.0331	.1468	.2541	.2787	.2187	.1239	.0467	.0092	.0004	.0000
	4	.0004	.0046	.0459	.1361	.2322	.2734	.2322	.1361	.0459	.0046	.0004
	5	.0000	.0004	.0092	.0467	.1239	.2187	.2787	.2541	.1468	.0331	.0054
	6	.0000	.0000	.0011	.0100	.0413	.1094	.2090	.2965	.2936	.1488	.0515
	7	.0000	.0000	.0001	.0012	.0079	.0312	.0896	.1977	.3355	.3826	.2793
	8	.0000	.0000	.0000	.0001	.0007	.0039	.0168	.0576	.1678	.4305	.6634

Appendix–C (Continued)

Table of Binomial Probabilities

n	x	p										
		.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
9	0	.6302	.3874	.1342	.0404	.0101	.0020	.0003	.0000	.0000	.0000	.0000
	1	.2985	.3874	.3020	.1556	.0605	.0176	.0035	.0004	.0000	.0000	.0000
	2	.0629	.1722	.3020	.2668	.1612	.0703	.0212	.0039	.0003	.0000	.0000
	3	.0077	.0446	.1762	.2668	.2508	.1641	.0743	.0210	.0028	.0001	.0000
	4	.0006	.0074	.0661	.1715	.2508	.2461	.1672	.0735	.0165	.0008	.0000
	5	.0000	.0008	.0165	.0735	.1672	.2461	.2508	.1715	.0661	.0074	.0006
	6	.0000	.0001	.0028	.0210	.0743	.1641	.2508	.2668	.1762	.0446	.0077
	7	.0000	.0000	.0003	.0039	.0212	.0703	.1612	.2668	.3020	.1722	.0629
	8	.0000	.0000	.0000	.0004	.0035	.0176	.0605	.1556	.3020	.3874	.2985
	9	.0000	.0000	.0000	.0000	.0003	.0020	.0101	.0404	.1342	.3874	.6302
10	0	.5987	.3487	.1074	.0282	.0060	.0010	.0001	.0000	.0000	.0000	.0000
	1	.3151	.3874	.2684	.1211	.0403	.0098	.0016	.0001	.0000	.0000	.0000
	2	.0746	.1937	.3020	.2335	.1209	.0439	.0106	.0014	.0001	.0000	.0000
	3	.0105	.0574	.2013	.2668	.2150	.1172	.0425	.0090	.0008	.0000	.0000
	4	.0010	.0112	.0881	.2001	.2508	.2051	.1115	.0368	.0055	.0001	.0000
	5	.0001	.0015	.0264	.1029	.2007	.2461	.2007	.1029	.0264	.0015	.0001
	6	.0000	.0001	.0055	.0368	.1115	.2051	.2508	.2001	.0881	.0112	.0010
	7	.0000	.0000	.0008	.0090	.0425	.1172	.2150	.2668	.2013	.0574	.0105
	8	.0000	.0000	.0001	.0014	.0106	.0439	.1209	.2335	.3020	.1937	.0746
	9	.0000	.0000	.0000	.0001	.0016	.0098	.0403	.1211	.2684	.3874	.3151
	10	.0000	.0000	.0000	.0000	.0001	.0010	.0060	.0282	.1074	.3487	.5987
11	0	.5688	.3138	.0859	.0198	.0036	.0005	.0000	.0000	.0000	.0000	.0000
	1	.3293	.3835	.2362	.0932	.0266	.0054	.0007	.0000	.0000	.0000	.0000
	2	.0867	.2131	.2953	.1998	.0887	.0269	.0052	.0005	.0000	.0000	.0000
	3	.0137	.0710	.2215	.2568	.1774	.0806	.0234	.0037	.0002	.0000	.0000
	4	.0014	.0158	.1107	.2201	.2365	.1611	.0701	.0173	.0017	.0000	.0000
	5	.0001	.0025	.0388	.1321	.2207	.2256	.1471	.0566	.0097	.0003	.0000
	6	.0000	.0003	.0097	.0566	.1471	.2256	.2207	.1321	.0388	.0025	.0001
	7	.0000	.0000	.0017	.0173	.0701	.1611	.2365	.2201	.1107	.0158	.0014
	8	.0000	.0000	.0002	.0037	.0234	.0806	.1774	.2568	.2215	.0710	.0137
	9	.0000	.0000	.0000	.0005	.0052	.0269	.0887	.1998	.2953	.2131	.0867
	10	.0000	.0000	.0000	.0000	.0007	.0054	.0266	.0932	.2362	.3835	.3293
	11	.0000	.0000	.0000	.0000	.0000	.0005	.0036	.0198	.0859	.3138	.5688
12	0	.5404	.2824	.0687	.0138	.0022	.0002	.0000	.0000	.0000	.0000	.0000
	1	.3413	.3766	.2062	.0712	.0174	.0029	.0003	.0000	.0000	.0000	.0000
	2	.0988	.2301	.2835	.1678	.0639	.0161	.0025	.0002	.0000	.0000	.0000
	3	.0173	.0852	.2362	.2397	.1419	.0537	.0125	.0015	.0001	.0000	.0000
	4	.0021	.0213	.1329	.2311	.2128	.1208	.0420	.0078	.0005	.0000	.0000
	5	.0002	.0038	.0532	.1585	.2270	.1934	.1009	.0291	.0033	.0000	.0000
	6	.0000	.0005	.0155	.0792	.1766	.2256	.1766	.0792	.0155	.0005	.0000
	7	.0000	.0000	.0033	.0291	.1009	.1934	.2270	.1585	.0532	.0038	.0002
	8	.0000	.0000	.0005	.0078	.0420	.1208	.2128	.2311	.1329	.0213	.0021
	9	.0000	.0000	.0001	.0015	.0125	.0537	.1419	.2397	.2362	.0852	.0173
	10	.0000	.0000	.0000	.0002	.0025	.0161	.0639	.1678	.2835	.2301	.0988
	11	.0000	.0000	.0000	.0000	.0003	.0029	.0174	.0712	.2062	.3766	.3413
	12	.0000	.0000	.0000	.0000	.0000	.0002	.0022	.0138	.0687	.2824	.5404

Appendix–C (Continued)

Table of Binomial Probabilities

		p										
n	*x*	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
13	0	.5133	.2542	.0550	.0097	.0013	.0001	.0000	.0000	.0000	.0000	.0000
	1	.3512	.3672	.1787	.0540	.0113	.0016	.0001	.0000	.0000	.0000	.0000
	2	.1109	.2448	.2680	.1388	.0453	.0095	.0012	.0001	.0000	.0000	.0000
	3	.0214	.0997	.2457	.2181	.1107	.0349	.0065	.0006	.0000	.0000	.0000
	4	.0028	.0277	.1535	.2337	.1845	.0873	.0243	.0034	.0001	.0000	.0000
	5	.0003	.0055	.0691	.1803	.2214	.1571	.0656	.0142	.0011	.0000	.0000
	6	.0000	.0008	.0230	.1030	.1968	.2095	.1312	.0442	.0058	.0001	.0000
	7	.0000	.0001	.0058	.0442	.1312	.2095	.1968	.1030	.0230	.0008	.0000
	8	.0000	.0000	.0011	.0142	.0656	.1571	.2214	.1803	.0691	.0055	.0003
	9	.0000	.0000	.0001	.0034	.0243	.0873	.1845	.2337	.1535	.0277	.0028
	10	.0000	.0000	.0000	.0006	.0065	.0349	.1107	.2181	.2457	.0997	.0214
	11	.0000	.0000	.0000	.0001	.0012	.0095	.0453	.1388	.2680	.2448	.1109
	12	.0000	.0000	.0000	.0000	.0001	.0016	.0113	.0540	.1787	.3672	.3512
	13	.0000	.0000	.0000	.0000	.0000	.0001	.0013	.0097	.0550	.2542	.5133
14	0	.4877	.2288	.0440	.0068	.0008	.0001	.0000	.0000	.0000	.0000	.0000
	1	.3593	.3559	.1539	.0407	.0073	.0009	.0001	.0000	.0000	.0000	.0000
	2	.1229	.2570	.2501	.1134	.0317	.0056	.0005	.0000	.0000	.0000	.0000
	3	.0259	.1142	.2501	.1943	.0845	.0222	.0033	.0002	.0000	.0000	.0000
	4	.0037	.0349	.1720	.2290	.1549	.0611	.0136	.0014	.0000	.0000	.0000
	5	.0004	.0078	.0860	.1963	.2066	.1222	.0408	.0066	.0003	.0000	.0000
	6	.0000	.0013	.0322	.1262	.2066	.1833	.0918	.0232	.0020	.0000	.0000
	7	.0000	.0002	.0092	.0618	.1574	.2095	.1574	.0618	.0092	.0002	.0000
	8	.0000	.0000	.0020	.0232	.0918	.1833	.2066	.1262	.0322	.0013	.0000
	9	.0000	.0000	.0003	.0066	.0408	.1222	.2066	.1963	.0860	.0078	.0004
	10	.0000	.0000	.0000	.0014	.0136	.0611	.1549	.2290	.1720	.0349	.0037
	11	.0000	.0000	.0000	.0002	.0033	.0222	.0845	.1943	.2501	.1142	.0259
	12	.0000	.0000	.0000	.0000	.0005	.0056	.0317	.1134	.2501	.2570	.1229
	13	.0000	.0000	.0000	.0000	.0001	.0009	.0073	.0407	.1539	.3559	.3593
	14	.0000	.0000	.0000	.0000	.0000	.0001	.0008	.0068	.0440	.2288	.4877
15	0	.4633	.2059	.0352	.0047	.0005	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3658	.3432	.1319	.0305	.0047	.0005	.0000	.0000	.0000	.0000	.0000
	2	.1348	.2669	.2309	.0916	.0219	.0032	.0003	.0000	.0000	.0000	.0000
	3	.0307	.1285	.2501	.1700	.0634	.0139	.0016	.0001	.0000	.0000	.0000
	4	.0049	.0428	.1876	.2186	.1268	.0417	.0074	.0006	.0000	.0000	.0000
	5	.0006	.0105	.1032	.2061	.1859	.0916	.0245	.0030	.0001	.0000	.0000
	6	.0000	.0019	.0430	.1472	.2066	.1527	.0612	.0116	.0007	.0000	.0000
	7	.0000	.0003	.0138	.0811	.1771	.1964	.1181	.0348	.0035	.0000	.0000
	8	.0000	.0000	.0035	.0348	.1181	.1964	.1771	.0811	.0138	.0003	.0000
	9	.0000	.0000	.0007	.0116	.0612	.1527	.2066	.1472	.0430	.0019	.0000
	10	.0000	.0000	.0001	.0030	.0245	.0916	.1859	.2061	.1032	.0105	.0006
	11	.0000	.0000	.0000	.0006	.0074	.0417	.1268	.2186	.1876	.0428	.0049
	12	.0000	.0000	.0000	.0001	.0016	.0139	.0634	.1700	.2501	.1285	.0307
	13	.0000	.0000	.0000	.0000	.0003	.0032	.0219	.0916	.2309	.2669	.1348
	14	.0000	.0000	.0000	.0000	.0000	.0005	.0047	.0305	.1319	.3432	.3658
	15	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0047	.0352	.2059	.4633

Appendix–C (Continued)

Table of Binomial Probabilities

		p										
n	x	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
16	0	.4401	.1853	.0281	.0033	.0003	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3706	.3294	.1126	.0228	.0030	.0002	.0000	.0000	.0000	.0000	.0000
	2	.1463	.2745	.2111	.0732	.0150	.0018	.0001	.0000	.0000	.0000	.0000
	3	.0359	.1423	.2463	.1465	.0468	.0085	.0008	.0000	.0000	.0000	.0000
	4	.0061	.0514	.2001	.2040	.1014	.0278	.0040	.0002	.0000	.0000	.0000
	5	.0008	.0137	.1201	.2099	.1623	.0667	.0142	.0013	.0000	.0000	.0000
	6	.0001	.0028	.0550	.1649	.1983	.1222	.0392	.0056	.0002	.0000	.0000
	7	.0000	.0004	.0197	.1010	.1889	.1746	.0840	.0185	.0012	.0000	.0000
	8	.0000	.0001	.0055	.0487	.1417	.1964	.1417	.0487	.0055	.0001	.0000
	9	.0000	.0000	.0012	.0185	.0840	.1746	.1889	.1010	.0197	.0004	.0000
	10	.0000	.0000	.0002	.0056	.0392	.1222	.1983	.1649	.0550	.0028	.0001
	11	.0000	.0000	.0000	.0013	.0142	.0666	.1623	.2099	.1201	.0137	.0008
	12	.0000	.0000	.0000	.0002	.0040	.0278	.1014	.2040	.2001	.0514	.0061
	13	.0000	.0000	.0000	.0000	.0008	.0085	.0468	.1465	.2463	.1423	.0359
	14	.0000	.0000	.0000	.0000	.0001	.0018	.0150	.0732	.2111	.2745	.1463
	15	.0000	.0000	.0000	.0000	.0000	.0002	.0030	.0228	.1126	.3294	.3706
	16	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0033	.0281	.1853	.4401
17	0	.4181	.1668	.0225	.0023	.0002	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3741	.3150	.0957	.0169	.0019	.0001	.0000	.0000	.0000	.0000	.0000
	2	.1575	.2800	.1914	.0581	.0102	.0010	.0001	.0000	.0000	.0000	.0000
	3	.0415	.1556	.2393	.1245	.0341	.0052	.0004	.0000	.0000	.0000	.0000
	4	.0076	.0605	.2093	.1868	.0796	.0182	.0021	.0001	.0000	.0000	.0000
	5	.0010	.0175	.1361	.2081	.1379	.0472	.0081	.0006	.0000	.0000	.0000
	6	.0001	.0039	.0680	.1784	.1839	.0944	.0242	.0026	.0001	.0000	.0000
	7	.0000	.0007	.0267	.1201	.1927	.1484	.0571	.0095	.0004	.0000	.0000
	8	.0000	.0001	.0084	.0644	.1606	.1855	.1070	.0276	.0021	.0000	.0000
	9	.0000	.0000	.0021	.0276	.1070	.1855	.1606	.0644	.0084	.0001	.0000
	10	.0000	.0000	.0004	.0095	.0571	.1484	.1927	.1201	.0267	.0007	.0000
	11	.0000	.0000	.0001	.0026	.0242	.0944	.1839	.1784	.0680	.0039	.0001
	12	.0000	.0000	.0000	.0006	.0081	.0472	.1379	.2081	.1361	.0175	.0010
	13	.0000	.0000	.0000	.0001	.0021	.0182	.0796	.1868	.2093	.0605	.0076
	14	.0000	.0000	.0000	.0000	.0004	.0052	.0341	.1245	.2393	.1556	.0415
	15	.0000	.0000	.0000	.0000	.0001	.0010	.0102	.0581	.1914	.2800	.1575
	16	.0000	.0000	.0000	.0000	.0000	.0001	.0019	.0169	.0957	.3150	.3741
	17	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0023	.0225	.1668	.4181

Appendix–C (Continued)

Table of Binomial Probabilities

		p										
n	*x*	**.05**	**.10**	**.20**	**.30**	**.40**	**.50**	**.60**	**.70**	**.80**	**.90**	**.95**
18	0	.3972	.1501	.0180	.0016	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3763	.3002	.0811	.0126	.0012	.0001	.0000	.0000	.0000	.0000	.0000
	2	.1683	.2835	.1723	.0458	.0069	.0006	.0000	.0000	.0000	.0000	.0000
	3	.0473	.1680	.2297	.1046	.0246	.0031	.0002	.0000	.0000	.0000	.0000
	4	.0093	.0700	.2153	.1681	.0614	.0117	.0011	.0000	.0000	.0000	.0000
	5	.0014	.0218	.1507	.2017	.1146	.0327	.0045	.0002	.0000	.0000	.0000
	6	.0002	.0052	.0816	.1873	.1655	.0708	.0145	.0012	.0000	.0000	.0000
	7	.0000	.0010	.0350	.1376	.1892	.1214	.0374	.0046	.0001	.0000	.0000
	8	.0000	.0002	.0120	.0811	.1734	.1669	.0771	.0149	.0008	.0000	.0000
	9	.0000	.0000	.0033	.0386	.1284	.1855	.1284	.0386	.0033	.0000	.0000
	10	.0000	.0000	.0008	.0149	.0771	.1669	.1734	.0811	.0120	.0002	.0000
	11	.0000	.0000	.0001	.0046	.0374	.1214	.1892	.1376	.0350	.0010	.0000
	12	.0000	.0000	.0000	.0012	.0145	.0708	.1655	.1873	.0816	.0052	.0002
	13	.0000	.0000	.0000	.0002	.0045	.0327	.1146	.2017	.1507	.0218	.0014
	14	.0000	.0000	.0000	.0000	.0011	.0117	.0614	.1681	.2153	.0700	.0093
	15	.0000	.0000	.0000	.0000	.0002	.0031	0246	.1046	.2297	.1680	.0473
	16	.0000	.0000	.0000	.0000	.0000	.0006	.0069	.0458	.1723	.2835	.1683
	17	.0000	.0000	.0000	.0000	.0000	.0001	.0012	.0126	.0811	.3002	.3763
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0016	.0180	.1501	.3972
19	0	.3774	.1351	.0144	.0011	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3774	.2852	.0685	.0093	.0008	.0000	.0000	.0000	.0000	.0000	.0000
	2	.1787	.2852	.1540	.0358	.0046	.0003	.0000	.0000	.0000	.0000	.0000
	3	.0533	.1796	.2182	.0869	.0175	.0018	.0001	.0000	.0000	.0000	.0000
	4	.0112	.0798	.2182	.1491	.0467	.0074	.0005	.0000	.0000	.0000	.0000
	5	.0018	.0266	.1636	.1916	.0933	.0222	.0024	.0001	.0000	.0000	.0000
	6	.0002	.0069	.0955	.1916	.1451	.0518	.0085	.0005	.0000	.0000	.0000
	7	.0000	.0014	.0443	.1525	.1797	.0961	.0237	.0022	.0000	.0000	.0000
	8	.0000	.0002	.0166	.0981	.1797	.1442	.0532	.0077	.0003	.0000	.0000
	9	.0000	.0000	.0051	.0514	.1464	.1762	.0976	.0220	.0013	.0000	.0000
	10	.0000	.0000	.0013	.0220	.0976	.1762	.1464	.0514	.0051	.0000	.0000
	11	.0000	.0000	.0003	.0077	.0532	.1442	.1797	.0981	.0166	.0002	.0000
	12	.0000	.0000	.0000	.0022	.0237	.0961	.1797	.1525	.0443	.0014	.0000
	13	.0000	.0000	.0000	.0005	.0085	.0518	.1451	.1916	.0955	.0069	.0002
	14	.0000	.0000	.0000	.0001	.0024	.0222	.0933	.1916	.1636	.0266	.0018
	15	.0000	.0000	.0000	.0000	.0005	.0074	.0467	.1491	.2182	.0798	.0112
	16	.0000	.0000	.0000	.0000	.0001	.0018	.0175	.0869	.2182	.1796	.0533
	17	.0000	.0000	.0000	.0000	.0000	.0003	.0046	.0358	.1540	.2852	.1787
	18	.0000	.0000	.0000	.0000	.0000	.0000	.0008	.0093	.0685	.2852	.3774
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0011	.0144	.1351	.3774

Appendix–C (Continued)

Table of Binomial Probabilities

		p										
n	*x*	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
20	0	.3585	.1216	.0115	.0008	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3774	.2702	.0576	.0068	.0005	.0000	.0000	.0000	.0000	.0000	.0000
	2	.1887	.2852	.1369	.0278	.0031	.0002	.0000	.0000	.0000	.0000	.0000
	3	.0596	.1901	.2054	.0716	.0123	.0011	.0000	.0000	.0000	.0000	.0000
	4	.0133	.0898	.2182	.1304	.0350	.0046	.0003	.0000	.0000	.0000	.0000
	5	.0022	.0319	.1746	.1789	.0746	.0148	.0013	.0000	.0000	.0000	.0000
	6	.0003	.0089	.1091	.1916	.1244	.0370	.0049	.0002	.0000	.0000	.0000
	7	.0000	.0020	.0545	.1643	.1659	.0739	.0146	.0010	.0000	.0000	.0000
	8	.0000	.0004	.0222	.1144	.1797	.1201	.0355	.0039	.0001	.0000	.0000
	9	.0000	.0001	.0074	.0654	.1597	.1602	.0710	.0120	.0005	.0000	.0000
	10	.0000	.0000	.0020	.0308	.1171	.1762	.1171	.0308	.0020	.0000	.0000
	11	.0000	.0000	.0005	.0120	.0710	.1602	.1597	.0654	.0074	.0001	.0000
	12	.0000	.0000	.0001	.0039	.0355	.1201	.1797	.1144	.0222	.0004	.0000
	13	.0000	.0000	.0000	.0010	.0146	.0739	.1659	.1643	.0545	.0020	.0000
	14	.0000	.0000	.0000	.0002	.0049	.0370	.1244	.1916	.1091	.0089	.0003
	15	.0000	.0000	.0000	.0000	.0013	.0148	.0746	.1789	.1746	.0319	.0022
	16	.0000	.0000	.0000	.0000	.0003	.0046	.0350	.1304	.2182	.0898	.0133
	17	.0000	.0000	.0000	.0000	.0000	.0011	.0123	.0716	.2054	.1901	.0596
	18	.0000	.0000	.0000	.0000	.0000	.0002	.0031	.0278	.1369	.2852	.1887
	19	.0000	.0000	.0000	.0000	.0000	.0000	.0005	.0068	.0576	.2702	.3774
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0008	.0115	.1216	.3585
21	0	.3406	.1094	.0092	.0006	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3764	.2553	.0484	.0050	.0003	.0000	.0000	.0000	.0000	.0000	.0000
	2	.1981	.2837	.1211	.0215	.0020	.0001	.0000	.0000	.0000	.0000	.0000
	3	.0660	.1996	.1917	.0585	.0086	.0006	.0000	.0000	.0000	.0000	.0000
	4	.0156	.0998	.2156	.1128	.0259	.0029	.0001	.0000	.0000	.0000	.0000
	5	.0028	.0377	.1833	.1643	.0588	.0097	.0007	.0000	.0000	.0000	.0000
	6	.0004	.0112	.1222	.1878	.1045	.0259	.0027	.0001	.0000	.0000	.0000
	7	.0000	.0027	.0655	.1725	.1493	.0554	.0087	.0005	.0000	.0000	.0000
	8	.0000	.0005	.0286	.1294	.1742	.0970	.0229	.0019	.0000	.0000	.0000
	9	.0000	.0001	.0103	.0801	.1677	.1402	.0497	.0063	.0002	.0000	.0000
	10	.0000	.0000	.0031	.0412	.1342	.1682	.0895	.0176	.0008	.0000	.0000
	11	.0000	.0000	.0008	.0176	.0895	.1682	.1342	.0412	.0031	.0000	.0000
	12	.0000	.0000	.0002	.0063	.0497	.1402	.1677	.0801	.0103	.0001	.0000
	13	.0000	.0000	.0000	.0019	.0229	.0970	.1742	.1294	.0286	.0005	.0000
	14	.0000	.0000	.0000	.0005	.0087	.0554	.1493	.1725	.0655	.0027	.0000
	15	.0000	.0000	.0000	.0001	.0027	.0259	.1045	.1878	.1222	.0112	.0004
	16	.0000	.0000	.0000	.0000	.0007	.0097	.0588	.1643	.1833	.0377	.0028
	17	.0000	.0000	.0000	.0000	.0001	.0029	.0259	.1128	.2156	.0998	.0156
	18	.0000	.0000	.0000	.0000	.0000	.0006	.0086	.0585	.1917	.1996	.0660
	19	.0000	.0000	.0000	.0000	.0000	.0001	.0020	.0215	.1211	.2837	.1981
	20	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0050	.0484	.2553	.3764
	21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0006	.0092	.1094	.3406

Appendix–C (Continued)

Table of Binomial Probabilities

		p										
n	*x*	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
22	0	.3235	.0985	.0074	.0004	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3746	.2407	.0406	.0037	.0002	.0000	.0000	.0000	.0000	.0000	.0000
	2	.2070	.2808	.1065	.0166	.0014	.0001	.0000	.0000	.0000	.0000	.0000
	3	.0726	.2080	.1775	.0474	.0060	.0004	.0000	.0000	.0000	.0000	.0000
	4	.0182	.1098	.2108	.0965	.0190	.0017	.0001	.0000	.0000	.0000	.0000
	5	.0034	.0439	.1898	.1489	.0456	.0063	.0004	.0000	.0000	.0000	.0000
	6	.0005	.0138	.1344	.1808	.0862	.0178	.0015	.0000	.0000	.0000	.0000
	7	.0001	.0035	.0768	.1771	.1314	.0407	.0051	.0002	.0000	.0000	.0000
	8	.0000	.0007	.0360	.1423	.1642	.0762	.0144	.0009	.0000	.0000	.0000
	9	.0000	.0001	.0140	.0949	.1703	.1186	.0336	.0032	.0001	.0000	.0000
	10	.0000	.0000	.0046	.0529	.1476	.1542	.0656	.0097	.0003	.0000	.0000
	11	.0000	.0000	.0012	.0247	.1073	.1682	.1073	.0247	.0012	.0000	.0000
	12	.0000	.0000	.0003	.0097	.0656	.1542	.1476	.0529	.0046	.0000	.0000
	13	.0000	.0000	.0001	.0032	.0336	.1186	.1703	.0949	.0140	.0001	.0000
	14	.0000	.0000	.0000	.0009	.0144	.0762	.1642	.1423	.0360	.0007	.0000
	15	.0000	.0000	.0000	.0002	.0051	.0407	.1314	.1771	.0768	.0035	.0001
	16	.0000	.0000	.0000	.0000	.0015	.0178	.0862	.1808	.1344	.0138	.0005
	17	.0000	.0000	.0000	.0000	.0004	.0063	.0456	.1489	.1898	.0439	.0034
	18	.0000	.0000	.0000	.0000	.0001	.0017	.0190	.0965	.2108	.1098	.0182
	19	.0000	.0000	.0000	.0000	.0000	.0004	.0060	.0474	.1775	.2080	.0726
	20	.0000	.0000	.0000	.0000	.0000	.0001	.0014	.0166	.1065	.2808	.2070
	21	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0037	.0406	.2407	.3746
	22	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0004	.0074	.0985	.3235
23	0	.3074	.0886	.0059	.0003	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3721	.2265	.0339	.0027	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	2	.2154	.2768	.0933	.0127	.0009	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0794	.2153	.1633	.0382	.0041	.0002	.0000	.0000	.0000	.0000	.0000
	4	.0209	.1196	.2042	.0818	.0138	.0011	.0000	.0000	.0000	.0000	.0000
	5	.0042	.0505	.1940	.1332	.0350	.0040	.0002	.0000	.0000	.0000	.0000
	6	.0007	.0168	.1455	.1712	.0700	.0120	.0008	.0000	.0000	.0000	.0000
	7	.0001	.0045	.0883	.1782	.1133	.0292	.0029	.0001	.0000	.0000	.0000
	8	.0000	.0010	.0442	.1527	.1511	.0584	.0088	.0004	.0000	.0000	.0000
	9	.0000	.0002	.0184	.1091	.1679	.0974	.0221	.0016	.0000	.0000	.0000
	10	.0000	.0000	.0064	.0655	.1567	.1364	.0464	.0052	.0001	.0000	.0000
	11	.0000	.0000	.0019	.0332	.1234	.1612	.0823	.0142	.0005	.0000	.0000
	12	.0000	.0000	.0005	.0142	.0823	.1612	.1234	.0332	.0019	.0000	.0000
	13	.0000	.0000	.0001	.0052	.0464	.1364	.1567	.0655	.0064	.0000	.0000
	14	.0000	.0000	.0000	.0016	.0221	.0974	.1679	.1091	.0184	.0002	.0000
	15	.0000	.0000	.0000	.0004	.0088	.0584	.1511	.1527	.0442	.0010	.0000
	16	.0000	.0000	.0000	.0001	.0029	.0292	.1133	.1782	.0883	.0045	.0001
	17	.0000	.0000	.0000	.0000	.0008	.0120	.0700	.1712	.1455	.0168	.0007
	18	.0000	.0000	.0000	.0000	.0002	.0040	.0350	.1332	.1940	.0505	.0042
	19	.0000	.0000	.0000	.0000	.0000	.0011	.0138	.0818	.2042	.1196	.0209
	20	.0000	.0000	.0000	.0000	.0000	.0002	.0041	.0382	.1633	.2153	.0794
	21	.0000	.0000	.0000	.0000	.0000	.0000	.0009	.0127	.0933	.2768	.2154
	22	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0027	.0339	.2265	.3721
	23	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0003	.0059	.0886	.3074

Appendix–C (Continued)

Table of Binomial Probabilities

		p										
n	*x*	.05	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
24	0	.2920	.0798	.0047	.0002	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3688	.2127	.0283	.0020	.0001	.0000	.0000	.0000	.0000	.0000	.0000
	2	.2232	.2718	.0815	.0097	.0006	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0862	.2215	.1493	.0305	.0028	.0001	.0000	.0000	.0000	.0000	.0000
	4	.0238	.1292	.1960	.0687	.0099	.0006	.0000	.0000	.0000	.0000	.0000
	5	.0050	.0574	.1960	.1177	.0265	.0025	.0001	.0000	.0000	.0000	.0000
	6	.0008	.0202	.1552	.1598	.0560	.0080	.0004	.0000	.0000	.0000	.0000
	7	.0001	.0058	.0998	.1761	.0960	.0206	.0017	.0000	.0000	.0000	.0000
	8	.0000	.0014	.0530	.1604	.1360	.0438	.0053	.0002	.0000	.0000	.0000
	9	.0000	.0003	.0236	.1222	.1612	.0779	.0141	.0008	.0000	.0000	.0000
	10	.0000	.0000	.0088	.0785	.1612	.1169	.0318	.0026	.0000	.0000	.0000
	11	.0000	.0000	.0028	.0428	.1367	.1488	.0608	.0079	.0002	.0000	.0000
	12	.0000	.0000	.0008	.0199	.0988	.1612	.0988	.0199	.0008	.0000	.0000
	13	.0000	.0000	.0002	.0079	.0608	.1488	.1367	.0428	.0028	.0000	.0000
	14	.0000	.0000	.0000	.0026	.0318	.1169	.1612	.0785	.0088	.0000	.0000
	15	.0000	.0000	.0000	.0008	.0141	.0779	.1612	.1222	.0236	.0003	.0000
	16	.0000	.0000	.0000	.0002	.0053	.0438	.1360	.1604	.0530	.0014	.0000
	17	.0000	.0000	.0000	.0000	.0017	.0206	.0960	.1761	.0998	.0058	.0001
	18	.0000	.0000	.0000	.0000	.0004	.0080	.0560	.1598	.1552	.0202	.0008
	19	.0000	.0000	.0000	.0000	.0001	.0025	.0265	.1177	.1960	.0574	.0050
	20	.0000	.0000	.0000	.0000	.0000	.0006	.0099	.0687	.1960	.1292	.0238
	21	.0000	.0000	.0000	.0000	.0000	.0001	.0028	.0305	.1493	.2215	.0862
	22	.0000	.0000	.0000	.0000	.0000	.0000	.0006	.0097	.0815	.2718	.2232
	23	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0020	.0283	.2127	.3688
	24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0002	.0047	.0798	.2920
25	0	.2774	.0718	.0038	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	1	.3650	.1994	.0236	.0014	.0000	.0000	.0000	.0000	.0000	.0000	.0000
	2	.2305	.2659	.0708	.0074	.0004	.0000	.0000	.0000	.0000	.0000	.0000
	3	.0930	.2265	.1358	.0243	.0019	.0001	.0000	.0000	.0000	.0000	.0000
	4	.0269	.1384	.1867	.0572	.0071	.0004	.0000	.0000	.0000	.0000	.0000
	5	.0060	.0646	.1960	.1030	.0199	.0016	.0000	.0000	.0000	.0000	.0000
	6	.0010	.0239	.1633	.1472	.0442	.0053	.0002	.0000	.0000	.0000	.0000
	7	.0001	.0072	.1108	.1712	.0800	.0143	.0009	.0000	.0000	.0000	.0000
	8	.0000	.0018	.0623	.1651	.1200	.0322	.0031	.0001	.0000	.0000	.0000
	9	.0000	.0004	.0294	.1336	.1511	.0609	.0088	.0004	.0000	.0000	.0000
	10	.0000	.0001	.0118	.0916	.1612	.0974	.0212	.0013	.0000	.0000	.0000
	11	.0000	.0000	.0040	.0536	.1465	.1328	.0434	.0042	.0001	.0000	.0000
	12	.0000	.0000	.0012	.0268	.1140	.1550	.0760	.0115	.0003	.0000	.0000
	13	.0000	.0000	.0003	.0115	.0760	.1550	.1140	.0268	.0012	.0000	.0000
	14	.0000	.0000	.0001	.0042	.0434	.1328	.1465	.0536	.0040	.0000	.0000
	15	.0000	.0000	.0000	.0013	.0212	.0974	.1612	.0916	.0118	.0001	.0000
	16	.0000	.0000	.0000	.0004	.0088	.0609	.1511	.1336	.0294	.0004	.0000
	17	.0000	.0000	.0000	.0001	.0031	.0322	.1200	.1651	.0623	.0018	.0000
	18	.0000	.0000	.0000	.0000	.0009	.0143	.0800	.1712	.1108	.0072	.0001
	19	.0000	.0000	.0000	.0000	.0002	.0053	.0442	.1472	.1633	.0239	.0010
	20	.0000	.0000	.0000	.0000	.0000	.0016	.0199	.1030	.1960	.0646	.0060
	21	.0000	.0000	.0000	.0000	.0000	.0004	.0071	.0572	.1867	.1384	.0269
	22	.0000	.0000	.0000	.0000	.0000	.0001	.0019	.0243	.1358	.2265	.0930
	23	.0000	.0000	.0000	.0000	.0000	.0000	.0004	.0074	.0708	.2659	.2305
	24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0014	.0236	.1994	.3650
	25	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0038	.0718	.2774

Appendix–D

Cumulative Poisson Distributions

	λ									
x	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0	.9048	.8187	.7408	.6703	.6065	.5488	.4966	.4493	.4066	.3679
1	.0905	.1637	.2222	.2681	.3033	.3293	.3476	.3595	.3659	.3679
2	.0045	.0164	.0333	.0536	.0758	.0988	.1217	.1438	.1647	.1839
3	.0002	.0011	.0033	.0072	.0126	.0198	.0284	.0383	.0494	.0613
4	.0000	.0001	.0003	.0007	.0016	.0030	.0050	.0077	.0111	.0153
5	.0000	.0000	.0000	.0001	.0002	.0004	.0007	.0012	.0020	.0031
6	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0003	.0005
7	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

	λ									
x	**1.1**	**1.2**	**1.3**	**1.4**	**1.5**	**1.6**	**1.7**	**1.8**	**1.9**	**2.0**
0	.3329	.3012	.2725	.2466	.2231	.2019	.1827	.1653	.1496	.1353
1	.3662	.3614	.3543	.3452	.3347	.3230	.3106	.2975	.2842	.2707
2	.2014	.2169	.2303	.2417	.2510	.2584	.2640	.2678	.2700	.2707
3	.0738	.0867	.0998	.1128	.1255	.1378	.1496	.1607	.1710	.1804
4	.0203	.0260	.0324	.0395	.0471	.0551	.0636	.0723	.0812	.0902
5	.0045	.0062	.0084	.0111	.0141	.0176	.0216	.0260	.0309	.0361
6	.0008	.0012	.0018	.0026	.0035	.0047	.0061	.0078	.0098	.0120
7	.0001	.0002	.0003	.0005	.0008	.0011	.0015	.0020	.0027	.0034
8	.0000	.0000	.0001	.0001	.0001	.0002	.0003	.0005	.0006	.0009
9	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002

	λ									
x	**2.1**	**2.2**	**2.3**	**2.4**	**2.5**	**2.6**	**2.7**	**2.8**	**2.9**	**3.0**
0	.1225	.1108	.1003	.0907	.0821	.0743	.0672	.0608	0550	.0498
1	.2572	.2438	.2306	.2177	.2052	.1931	.1815	.1703	.1596	.1494
2	.2700	.2681	.2652	.2613	.2565	.2510	.2450	.2384	.2314	.2240
3	.1890	.1966	.2033	.2090	.2138	.2176	.2205	.2225	.2237	.2240
4	.0992	.1082	.1169	.1254	.1336	.1414	.1488	.1557	.1622	.1680
5	.0417	.0476	.0538	.0602	.0668	.0735	.0804	.0872	.0940	.1008
6	.0146	.0174	.0206	.0241	.0278	.0319	.0362	.0407	.0455	.0504
7	.0044	.0055	.0068	.0083	.0099	.0118	.0139	.0163	.0188	.0216
8	.0011	.0015	.0019	.0025	.0031	.0038	.0047	.0057	.0068	.0081
9	.0003	.0004	.0005	.0007	.0009	.0011	.0014	.0018	.0022	.0027
10	.0001	.0001	.0001	.0002	.0002	.0003	.0004	.0005	.0006	.0008
11	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0002	.0002
12	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

Appendix–D (Continued)

Cumulative Poisson Distributions

	λ									
x	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	.0450	.0408	.0369	.0334	.0302	.0273	.0247	.0224	.0202	.0183
1	.1397	.1304	.1217	.1135	.1057	.0984	.0915	.0850	.0789	.0733
2	.2165	.2087	.2008	.1929	.1850	.1771	.1692	.1615	.1539	.1465
3	.2237	.2226	.2209	.2186	.2158	.2125	.2087	.2046	.2001	.1954
4	.1733	.1781	.1823	.1858	.1888	.1912	.1931	.1944	.1951	.1954
5	.1075	.1140	.1203	.1264	.1322	.1377	.1429	.1477	.1522	.1563
6	.0555	.0608	.0662	.0716	.0771	.0826	.0881	.0936	.0989	.1042
7	.0246	.0278	.0312	.0348	.0385	.0425	.0466	.0508	.0551	.0595
8	.0095	.0111	.0129	.0148	.0169	.0191	.0215	.0241	.0269	.0298
9	.0033	.0040	.0047	.0056	.0066	.0076	.0089	.0102	.0116	.0132
10	.0010	.0013	.0016	.0019	.0023	.0028	.0033	.0039	.0045	.0053
11	.0003	.0004	.0005	.0006	.0007	.0009	.0011	.0013	.0016	.0019
12	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005	.0006
13	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0002	.0002
14	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

	λ									
x	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	.0166	.0150	.0136	.0123	.0111	.0101	.0091	.0082	.0074	.0067
1	.0679	.0630	.0583	.0540	.0500	.0462	.0427	.0395	.0365	.0337
2	.1393	.1323	.1254	.1188	.1125	.1063	.1005	.0948	.0894	.0842
3	.1904	.1852	.1798	.1743	.1687	.1631	.1574	.1517	.1460	.1404
4	.1951	.1944	.1933	.1917	.1898	.1875	.1849	.1820	.1789	.1755
5	.1600	.1633	.1662	.1687	.1708	.1725	.1738	.1747	.1753	.1755
6	.1039	.1143	.1191	.1237	.1281	.1323	.1362	.1398	.1432	.1462
7	.0640	.0686	.0732	.0778	.0824	.0869	.0914	.0959	.1002	.1044
8	.0328	.0360	.0393	0428	.0463	.0500	.0537	.0575	.0614	.0653
9	.0150	.0168	.0188	.0209	.0232	.0255	.0281	.0307	.0334	.0363
10	.0061	.0071	.0081	.0092	.0104	.0118	.0132	.0147	.0164	.0181
11	.0023	.0027	.0032	.0037	.0043	.0049	.0056	.0064	.0073	.0082
12	.0008	.0009	.0011	.0014	.0016	.0019	.0022	.0026	.0030	.0034
13	.0002	.0003	.0004	.0005	.0006	.0007	.0008	.0009	.0011	.0013
14	.0001	.0001	.0001	.0001	.0002	.0002	.0003	.0003	.0004	.0005
15	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0002

Appendix–D (Continued)

Cumulative Poisson Distributions

	λ									
x	**5.1**	**5.2**	**5.3**	**5.4**	**5.5**	**5.6**	**5.7**	**5.8**	**5.9**	**6.0**
0	.0061	.0055	.0050	.0045	.0041	.0037	.0033	.0030	.0027	.0025
1	.0311	.0287	.0265	.0244	.0225	.0207	.0191	.0176	.0162	.0149
2	.0793	.0746	.0701	.0659	.0618	.0580	.0544	.0509	.0477	.0446
3	.1348	.1293	.1239	.1185	.1133	.1082	.1033	.0985	.0938	.0892
4	.1719	.1681	.1641	.1600	.1558	.1515	.1472	.1428	.1383	.1339
5	.1753	.1748	.1740	.1728	.1714	.1697	.1678	.1656	.1632	.1606
6	.1490	.1515	.1537	.1555	.1571	.1584	.1594	.1601	.1605	.1606
7	.1086	.1125	.1163	.1200	.1234	.1267	.1298	.1326	.1353	.1377
8	.0692	.0731	.0771	.0810	.0849	.0887	.0925	.0962	.0998	.1033
9	.0392	.0423	.0454	.0486	.0519	.0552	0586	.0620	0.654	.0688
10	.0200	.0220	.0241	.0262	.0285	.0309	.0334	.0359	.0386	.0413
11	.0093	.0104	.0116	.0129	.0143	.0157	.0173	.0190	.0207	.0225
12	.0039	.0045	.0051	.0058	.0065	.0073	.0082	.0092	.0102	.0113
13	.0015	.0018	.0021	.0024	.0028	.0032	.0036	.0041	.0046	.0052
14	.0006	.0007	.0008	.0009	.0011	.0013	.0015	.0017	.0019	.0022
15	.0002	.0002	.0003	.0003	.0004	.0005	.0006	.0007	.0008	.0009
16	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003
17	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001

	λ									
x	**6.1**	**6.2**	**6.3**	**6.4**	**6.5**	**6.6**	**6.7**	**6.8**	**6.9**	**7.0**
0	.0022	.0020	.0018	.0017	.0015	.0014	.0012	.0011	.0010	.0009
1	.0137	.0126	.0116	.0106	.0098	.0090	.0082	.0076	.0070	.0064
2	.0417	.0390	.0364	.0340	.0318	.0296	.0276	.0258	.0240	.0223
3	.0848	.0806	.0765	.0726	.0688	.0652	.0617	.0584	.0552	.0521
4	.1294	.1249	.1205	.1162	.1118	.1076	.1034	.0992	.0952	.0912
5	.1579	.1549	.1519	.1487	.1454	.1420	.1385	.1349	.1314	.1277
6	.1605	.1601	.1595	.1586	.1575	.1562	.1546	.1529	.1511	.1490
7	.1399	.1418	.1435	.1450	.1462	.1472	.1480	.1486	.1489	.1490
8	.1066	.1099	.1130	.1160	.1188	.1215	.1240	.1263	.1284	.1304
9	.0723	.0757	.0791	.0825	.0858	.0891	.0923	.0954	.0985	.1014
10	.0441	.0469	.0498	.0528	.0558	.0588	.0618	.069	.0679	.0710
11	.0244	.0265	.0285	.0307	.0330	.0353	.0377	.0401	.0426	.0452
12	.0124	.0137	.0150	.0164	.0179	.0194	.0210	.0227	.0245	.0263
13	.0058	.0065	.0073	.0081	.0089	.0099	.0108	.0119	.0130	.0142
14	.0025	.0029	.0033	.0037	.0041	.0046	.0052	.0058	.0064	.0071
15	.0010	.0012	.0014	.0016	.0018	.0020	.0023	.0026	.0029	.0033
16	.0004	.0005	.0005	.0006	.0007	.0008	.0010	.0011	.0013	.0014
17	.0001	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0006
18	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002
19	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001

Appendix–D (Continued)

Cumulative Poisson Distributions

	λ									
x	**7.1**	**7.2**	**7.3**	**7.4**	**7.5**	**7.6**	**7.7**	**7.8**	**7.9**	**8.0**
0	.0008	.0007	.0007	.0006	.0006	.0005	.0005	.0004	.0004	.0003
1	.0059	.0054	.0049	.0045	.0041	.0038	.0035	.0032	.0029	.0027
2	.0208	.0194	.0180	.0167	.0156	.0145	.0134	.0125	.0116	.0107
3	.0492	.0464	.0438	.0413	.0389	.0366	.0345	.0324	.0305	.0286
4	.0874	.0836	.0799	.0764	.0729	.0696	.0663	.0632	.0602	.0573
5	.1241	.1204	.1167	.1130	.1094	.1057	.1021	.0986	.0951	.0916
6	.1468	.1445	.1420	.1394	.1367	.1339	.1311	.1282	.1252	.1221
7	.1489	.1486	.1481	.1474	.1465	.1454	.1442	.1428	.1413	.1396
8	.1321	.1337	.1351	.1363	.1373	.1381	.1388	.1392	.1395	.1396
9	.1042	.1070	.1096	.1121	.1144	.1167	.1187	.1207	.1224	.1241
10	.0740	.0770	.0800	.0829	.0858	.0887	.0914	.0941	.0967	.0993
11	.0478	.0504	.0531	.0558	.0585	.0613	.0640	.0667	.0695	.0722
12	.0283	.0303	.0323	.0344	.0366	.0388	.0411	.0434	.0457	.0481
13	.0154	.0168	.0181	.0196	.0211	.0227	.0243	.0260	.0278	.0296
14	.0078	.0086	.0095	.0104	.0113	.0123	.0134	.0145	.0157	.0169
15	.0037	.0041	.0046	.0051	.0057	.0062	.0069	.0075	.0083	.0090
16	.0016	.0019	.0021	.0024	.0026	.0030	.0033	.0037	.0041	.0045
17	.0007	.0008	.0009	.0010	.0012	.0013	.0015	.0017	.0019	.0021
18	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
19	.0001	.0001	.0001	.0002	.0002	.0002	.0003	.0003	.0003	.0004
20	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002
21	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001

	λ									
x	**8.1**	**8.2**	**8.3**	**8.4**	**8.5**	**8.6**	**8.7**	**8.8**	**8.9**	**9.0**
0	.0003	.0003	.0002	.0002	.0002	.0002	.0002	.0002	.0001	.0001
1	.0025	.0023	.0021	.0019	.0017	.0016	.0014	.0013	.0012	.0011
2	.0100	.0092	.0086	.0079	.0074	.0068	.0063	.0058	.0054	.0050
3	.0269	.0252	.0237	.0222	.0208	.0195	.0183	.0171	.0160	.0150
4	.0544	.0517	.0491	.0466	.0443	.0420	.0398	.0377	.0357	.0337
5	.0882	.0849	.0816	.0784	.0752	.0722	.0692	.0663	.0635	.0607
6	.1191	.1160	.1128	.1097	.1066	.1034	.1003	.0972	.0941	.0911
7	.1378	.1358	.1338	.1317	.1294	.1271	.1247	.1222	.1197	.1171
8	.1395	.1392	.1388	.1382	.1375	.1366	.1356	.1344	.1332	.1318
9	.1255	.1269	.1280	.1290	.1299	.1306	.1311	.1315	.1317	.1318
10	.1017	.1040	.1063	.1084	.1104	.1123	.1140	.1157	.1172	.1186
11	.0749	.0775	.0802	.0828	.0853	.0878	.0902	.0925	.0948	.0970
12	.0505	.0530	.0555	.0579	.0604	.0629	.0654	.0679	.0703	.0728
13	.0315	.0334	.0354	.0374	.0395	.0416	.0438	.0459	.0481	.0504
14	.0182	.0196	.0210	.0225	.0240	.0256	.0272	.0289	.0306	.0324
15	.0098	.0107	.0116	.0126	.0136	.0147	.0158	.0169	.0182	.0194
16	.0050	.0055	.0060	.0066	.0072	.0079	.0086	.0093	.0101	.0109
17	.0024	.0026	.0029	.0033	.0036	.0040	.0044	.0048	.0053	.0058
18	.0011	.0012	.0014	.0015	.0017	.0019	.0021	.0024	.0026	.0029
19	.0005	.0005	.0006	.0007	.0008	.0009	.0010	.0011	.0012	.0014
20	.0002	.0002	.0002	.0003	.0003	.0004	.0004	.0005	.0005	.0006
21	.0001	.0001	.0001	.0001	.0001	.0002	.0002	.0002	.0002	.0003
22	.0000	.0000	.0000	.0000	.0001	.0001	.0001	.0001	.0001	.0001

Appendix–D (Continued)

Cumulative Poisson Distributions

	λ									
x	**9.1**	**9.2**	**9.3**	**9.4**	**9.5**	**9.6**	**9.7**	**9.8**	**9.9**	**10**
0	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0000
1	.0010	.0009	.0009	.0008	.0007	.0007	.0006	.0005	.0005	.0005
2	.0046	.0043	.0040	.0037	.0034	.0031	.0029	.0027	.0025	.0023
3	.0140	.0131	.0123	.0115	.0107	.0100	.0093	.0087	.0081	.0076
4	.0319	.0302	.0285	.0269	.0254	.0240	.0226	.0213	.0201	.0189
5	.0581	.0555	.0530	.0506	.0483	.0460	.0439	.0418	.0398	.0378
6	.0881	.0851	.0822	.0793	.0764	.0736	.0709	.0682	.0656	.0631
7	.1145	.1118	.1091	.1064	.1037	.1010	.0982	.0955	.0928	.0901
8	.1302	.1286	.1269	.1251	.1232	.1212	.1191	.1170	.1148	.1126
9	.1317	.1315	.1311	.1306	.1300	.1293	.1284	.1274	.1263	.1251
10	.1198	.1209	.1219	.1228	.1235	.1241	.1245	.1249	.1250	.1251
11	.0991	.1012	.1031	.1049	.1067	.1083	.1098	.1112	.1125	.1137
12	.0752	.0776	.0799	.0822	.0844	.0866	.0888	.0908	.0928	.0948
13	.0526	.0549	.0572	.0594	.0617	.0640	.0662	.0685	.0707	.0729
14	.0342	.0361	.0380	.0399	.0419	.0439	.0459	.0479	.0500	.0521
15	.0208	.0221	.0235	.0250	.0265	.0281	.0297	.0313	.0330	.0347
16	.0118	.0127	.0137	.0147	.0157	.0168	.0180	.0192	.0204	.0217
17	.0063	.0069	.0075	.0081	.0088	.0095	.0103	.0111	.0119	.0128
18	.0032	.0035	.0039	.0042	.0046	.0051	.0055	.0060	.0065	.0071
19	.0015	.0017	.0019	.0021	.0023	.0026	.0028	.0031	.0034	.0037
20	.0007	.0008	.0009	.0010	.0011	.0012	.0014	.0015	.0017	.0019
21	.0003	.0003	.0004	.0004	.0005	.0006	.0006	.0007	.0008	.0009
22	.0001	.0001	.0002	.0002	.0002	.0002	.0003	.0003	.0004	.0004
23	.0000	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002
24	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0001	.0001

Appendix–D (Continued)

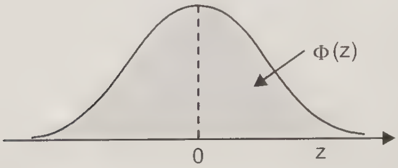
Cumulative Poisson Distributions

	λ									
x	11	12	13	14	15	16	17	18	19	20
0	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
1	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
2	.0010	.0004	.0002	.0001	.0000	.0000	.0000	.0000	.0000	.0000
3	.0037	.0018	.0008	.0004	.0002	.0001	.0000	.0000	.0000	.0000
4	.0102	.0053	.0027	.0013	.0006	.0003	.0001	.0001	.0000	.0000
5	.0224	.0127	.0070	.0037	.0019	.0010	.0005	.0002	.0001	.0001
6	.0411	.0255	.0152	.0087	.0048	.0026	.0014	.0007	.0004	.0002
7	.0646	.0437	.0281	.0174	.0104	.0060	.0034	.0019	.0010	.0005
8	.0888	.0655	.0457	.0304	.0194	.0120	.0072	.0042	.0024	.0013
9	.1085	.0874	.0661	.0473	.0324	.0213	.0135	.0083	.0050	.0029
10	.1194	.1048	.0859	.0663	.0486	.0341	.0230	.0150	.0095	.0058
11	.1194	.1144	.1015	.0844	.0663	.0496	.0355	.0245	.0164	.0106
12	.1094	.1144	.1099	.0984	.0829	.0661	.0504	.0368	.0259	.0176
13	.0926	.1056	.1099	.1060	.0956	.0814	.0658	.0509	.0378	.0271
14	.0728	.0905	.1021	.1060	.1024	.0930	.0800	.0655	.0514	.0387
15	.0534	.0724	.0885	.0989	.1024	.0992	.0906	.0786	.0650	.0516
16	.0367	.0543	.0719	.0866	.0960	.0992	.0963	.0884	.0772	.0646
17	.0237	.0383	.0550	.0713	.0847	.0934	.0963	.0936	.0863	.0760
18	.0145	.0255	.0397	.0554	.0706	.0830	.0909	.0936	.0911	.0844
19	.0084	.0161	.0272	.0409	.0557	.0699	.0814	.0887	.0911	.0888
20	.0046	.0097	.0177	.0286	.0418	.0559	.0692	.0798	.0866	.0888
21	.0024	.0055	.0109	.0191	.0299	.0426	.0560	.0684	.0783	.0846
22	.0012	.0030	.0065	.0121	.0204	.0310	.0433	.0560	.0676	.0769
23	.0006	.0016	.0037	.0074	.0133	.0216	.0320	.0438	.0559	.0669
24	.0003	.0008	.0020	.0043	.0083	.0144	.0226	.0328	.0442	.0557
25	.0001	.0004	.0010	.0024	.0050	.0092	.0154	.0237	.0336	.0446
26	.0000	.0002	.0005	.0013	.0029	.0057	.0101	.0164	.0246	.0343
27	.0000	.0001	.0002	.0007	.0016	.0034	.0063	.0109	.0173	.0254
28	.0000	.0000	.0001	.0003	.0009	.0019	.0038	.0070	.0117	.0181
29	.0000	.0000	.0001	.0002	.0004	.0011	.0023	.0044	.0077	.0125
30	.0000	.0000	.0000	.0001	.0002	.0006	.0013	.0026	.0049	.0083
31	.0000	.0000	.0000	.0000	.0001	.0003	.0007	.0015	.0030	.0054
32	.0000	.0000	.0000	.0000	.0001	.0001	.0004	.0009	.0018	.0034
33	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0005	.0010	.0020
34	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0006	.0012
35	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0003	.0007
36	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002	.0004
37	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001	.0002
38	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001
39	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0001

Appendix–E

Cumulative Standard Normal Distribution

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du$$

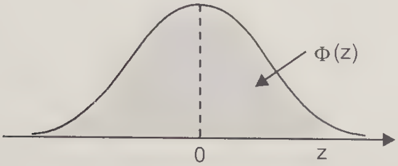


z	**–0.09**	**–0.08**	**–0.07**	**–0.06**	**–0.05**	**–0.04**	**–0.03**	**–0.02**	**–0.01**	**–0.00**	z
–3.9	0.000033	0.000034	0.000036	0.000037	0.000039	0.000041	0.000042	0.000044	0.000046	0.000048	–3.9
–3.8	0.000050	0.000052	0.000054	0.000057	0.000059	0.000062	0.000064	0.000067	0.000069	0.000072	–3.8
–3.7	0.000075	0.000078	0.000082	0.000085	0.000088	0.000092	0.000096	0.000100	0.000104	0.000108	–3.7
–3.6	0.000112	0.000117	0.000121	0.000126	0.000131	0.000136	0.000142	0.000147	0.000153	0.000159	–3.6
–3.5	0.000165	0.000172	0.000179	0.000185	0.000193	0.000200	0.000208	0.000216	0.000224	0.000233	–3.5
–3.4	0.000242	0.000251	0.000260	0.000270	0.000280	0.000291	0.000302	0.000313	0.000325	0.000337	–3.4
–3.3	0.000350	0.000362	0.000376	0.000390	0.000404	0.000419	0.000434	0.000450	0.000467	0.000483	–3.3
–3.2	0.000501	0.000519	0.000538	0.000557	0.000577	0.000598	0.000619	0.000641	0.000664	0.000687	–3.2
–3.1	0.000711	0.000736	0.000762	0.000789	0.000816	0.000845	0.000874	0.000904	0.000935	0.000968	–3.1
–3.0	0.001001	0.001035	0.001070	0.001107	0.001144	0.001183	0.001223	0.001264	0.001306	0.001350	–3.0
–2.9	0.001395	0.001441	0.001489	0.001538	0.001589	0.001641	0.001695	0.001750	0.001807	0.001866	–2.9
–2.8	0.001926	0.001988	0.002052	0.002118	0.002186	0.002256	0.002327	0.002401	0.002477	0.002555	–2.8
–2.7	0.002635	0.002718	0.002803	0.002890	0.002980	0.003072	0.003167	0.003264	0.003364	0.003467	–2.7
–2.6	0.003573	0.003681	0.003793	0.003907	0.004025	0.004145	0.004269	0.004396	0.004527	0.004661	–2.6
–2.5	0.004799	0.004940	0.005085	0.005234	0.005386	0.005543	0.005703	0.005868	0.006037	0.006210	–2.5
–2.4	0.006387	0.006569	0.006756	0.006947	0.007143	0.007344	0.007549	0.007760	0.007976	0.008198	–2.4
–2.3	0.008424	0.008656	0.008894	0.009137	0.009387	0.009642	0.009903	0.010170	0.010444	0.010724	–2.3
–2.2	0.011011	0.011304	0.011604	0.011911	0.012224	0.012545	0.012874	0.013209	0.013553	0.013903	–2.2
–2.1	0.014262	0.014629	0.015003	0.015386	0.015778	0.016177	0.016586	0.017003	0.017429	0.017864	–2.1
–2.0	0.018309	0.018763	0.019226	0.019699	0.020182	0.020675	0.021178	0.021692	0.022216	0.022750	–2.0
–1.9	0.023295	0.023852	0.024419	0.024998	0.025588	0.026190	0.026803	0.027429	0.028067	0.028717	–1.9
–1.8	0.029379	0.030054	0.030742	0.031443	0.032157	0.032884	0.033625	0.034379	0.035148	0.035930	–1.8
–1.7	0.036727	0.037538	0.038364	0.039204	0.040059	0.040929	0.041815	0.042716	0.043633	0.044565	–1.7
–1.6	0.045514	0.046479	0.047460	0.048457	0.049471	0.050503	0.050551	0.052616	0.053699	0054799	–1.6
–1.5	0.055917	0.057053	0.058208	0.059380	0.060571	0.061780	0.063008	0.064256	0.065522	0.066807	–1.5
–1.4	0.068112	0.069437	0.070781	0.072146	0.073529	0.074934	0.076359	0.077804	0.079270	0.080757	–1.4
–1.3	0.082264	0.083793	0.085343	0.086915	0.088508	0.090123	0.091759	0.093418	0.095098	0.096801	–1.3
–1.2	0.098525	0.100273	0.102042	0.103835	0.105650	0.107488	0.109349	0.111233	0.113140	0.115070	–1.2
–1.1	0.117023	0.119000	0.121001	0.123024	0.125072	0.127143	0.129238	0.131357	0.133500	0.135666	–1.1
–1.0	0.137857	0.140071	0.142310	0.144572	0.146859	0.149170	0.151505	0.153864	0.156248	0.158655	–1.0
–0.9	0.161087	0.163543	0.166023	0.168528	0.171056	0.173609	0.176185	0.178786	0.181411	0.184060	–0.9
–0.8	0.186733	0.189430	0.192150	0.194894	0.197662	0.200454	0.203269	0.206108	0.208970	0.211855	–0.8
–0.7	0.214764	0.217695	0.220650	0.223627	0.226627	0.229650	0.232695	0.235762	0.238852	0.241964	–0.7
–0.6	0.245097	0.248252	0.251429	0.254627	0.257846	0.261086	0.264347	0.267629	0.270931	0.274253	–0.6
–0.5	0.277595	0.280957	0.284339	0.287740	0.291160	0.294599	0.298056	0.301532	0.305026	0.308538	–0.5
–0.4	0.312067	0.315614	0.319178	0.322758	0.326355	0.329969	0.333598	0.337243	0.340903	0.344578	–0.4
–0.3	0.348268	0.351973	0.355691	0.359424	0.363169	0.366928	0.370700	0.374484	0.378281	0.382089	–0.3
–0.2	0.385908	0.389739	0.393580	0.397432	0.401294	0.405165	0.409046	0.412936	0.416834	0.420740	–0.2
–0.1	0.424655	0.428576	0.432505	0.436441	0.440382	0.444330	0.448283	0.452242	0.456205	0.460172	–0.1
0.0	0.464144	0.468119	0.472097	0.476078	0.480061	0.484047	0.488033	0.492022	0.496011	0.500000	0.0

Appendix–E (Continued)

Cumulative Standard Normal Distribution

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}\, du$$

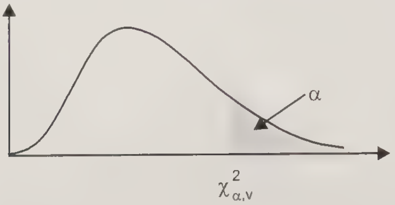


z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**	z
0.0	0.500000	0.503989	0.507978	0.511967	0.515953	0.519939	0.523922	0.527903	0.531881	0.535856	0.0
0.1	0.539828	0.543795	0.547758	0.551717	0.555760	0.559618	0.563559	0.567495	0.571424	0.575345	0.1
0.2	0.579260	0.583166	0.587064	0.590954	0.594835	0.598706	0.602568	0.606420	0.610261	0.614092	0.2
0.3	0.617911	0.621719	0.625516	0.629300	0.633072	0.636831	0.640576	0.644309	0.648027	0.651732	0.3
0.4	0.655422	0.659097	0.662757	0.666402	0.670031	0.673645	0.677242	0.680822	0.684386	0.687933	0.4
0.5	0.691462	0.694974	0.698468	0.701944	0.705401	0.708840	0.712260	0.715662	0.719043	0.722405	0.5
0.6	0.725747	0.729069	0.732371	0.735653	0.738914	0.742154	0.745373	0.748571	0.751748	0.754903	0.6
0.7	0.758036	0.761148	0.764238	0.767305	0.770350	0.773373	0.776373	0.779350	0.782305	0.785236	0.7
0.8	0.788145	0.791030	0.793892	0.796731	0.799546	0.802338	0.805106	0.807850	0.810570	0.813267	0.8
0.9	0.815940	0.818589	0.821214	0.823815	0.826391	0.828944	0.831472	0.833977	0.836457	0.838913	0.9
1.0	0.841345	0.843752	0.846136	0.848495	0.850830	0.853141	0.855428	0.857690	0.859929	0.862143	1.0
1.1	0.864334	0.866500	0.868643	0.870762	0.872857	0.874928	0.876976	0.878999	0.881000	0.882977	1.1
1.2	0.884930	0.886860	0.888767	0.890651	0.892512	0.894350	0.896165	0.897958	0.899727	0.901475	1.2
1.3	0.903199	0.904902	0.906582	0.908241	0.909877	0.911492	0.913085	0.914657	0.916207	0.917736	1.3
1.4	0.919243	0.920730	0.922196	0.923641	0.925066	0.926471	0.927855	0.929219	0.930563	0.931888	1.4
1.5	0.933193	0.934478	0.935744	0.936992	0.938220	0.939429	0.940620	0.941792	0.942947	0.944083	1.5
1.6	0.945201	0.946301	0.947384	0.948449	0.949497	0.950529	0.951543	0.952540	0.953521	0.954486	1.6
1.7	0.955435	0.956367	0.957284	0.958185	0.959071	0.959941	0.960796	0.961636	0.962462	0.963273	1.7
1.8	0.964070	0.964852	0.965621	0.966375	0.967116	0.967843	0.968557	0.969258	0.969946	0.970621	1.8
1.9	0.971283	0.971933	0.972571	0.973197	0.973810	0.974412	0.975002	0.975581	0.976148	0.976705	1.9
2.0	0.977250	0.977784	0.978308	0.978822	0.979325	0.979818	0.980301	0.980774	0.981237	0.981691	2.0
2.1	0.982136	0.982571	0.982997	0.983414	0.983823	0.984222	0.984614	0.984997	0.985371	0.985738	2.1
2.2	0.986097	0.986447	0.986791	0.987126	0.987455	0.987776	0.988089	0.988396	0.988696	0.988989	2.2
2.3	0.989276	0.989556	0.989830	0.990097	0.990358	0.990613	0.990863	0.991106	0.991344	0.991576	2.3
2.4	0.991802	0.992024	0.992240	0.992451	0.992656	0.992857	0.993053	0.993244	0.993431	0.993613	2.4
2.5	0.993790	0.993963	0.994132	0.994297	0.994457	0.994614	0.994766	0.994915	0.995060	0.995201	2.5
2.6	0.995339	0.995473	0.995604	0.995731	0.995855	0.995975	0.996093	0.996207	0.996319	0.996427	2.6
2.7	0.996533	0.996636	0.996736	0.996833	0.996928	0.997020	0.997110	0.997197	0.997282	0.997365	2.7
2.8	0.997445	0.997523	0.997599	0.997673	0.997744	0.997814	0.997882	0.997948	0.998012	0.998074	2.8
2.9	0.998134	0.998193	0.998250	0.998305	0.998359	0.998411	0.998462	0.998511	0.998559	0.998605	2.9
3.0	0.998650	0.998694	0.998736	0.998777	0.998817	0.998856	0.998893	0.998930	0.998965	0.998999	3.0
3.1	0.999032	0.999065	0.999096	0.999126	0.999155	0.999184	0.999211	0.999238	0.999264	0.999289	3.1
3.2	0.999313	0.999336	0.999359	0.999381	0.999402	0.999423	0.999443	0.999462	0.999481	0.999499	3.2
3.3	0.999517	0.999533	0.999550	0.999566	0.999581	0.999596	0.999610	0.999624	0.999638	0.999650	3.3
3.4	0.999663	0.999675	0.999687	0.999698	0.999709	0.999720	0.999730	0.999740	0.999749	0.999758	3.4
3.5	0.999767	0.999776	0.999784	0.999792	0.999800	0.999807	0.999815	0.999821	0.999828	0.999835	3.5
3.6	0.999841	0.999847	0.999853	0.999858	0.999864	0.999869	0.999874	0.999879	0.999883	0.999888	3.6
3.7	0.999892	0.999896	0.999900	0.999904	0.999908	0.999912	0.999915	0.999918	0.999922	0.999925	3.7
3.8	0.999928	0.999931	0.999933	0.999936	0.999938	0.999941	0.999943	0.999946	0.999948	0.999950	3.8
3.9	0.999952	0.999954	0.999956	0.999958	0.999959	0.999961	0.999963	0.999964	0.999966	0.999967	3.9

Appendix–F

chi-square Distribution Table

Percentage Points $\chi^2_{\alpha v}$ of the chi-square Distribution

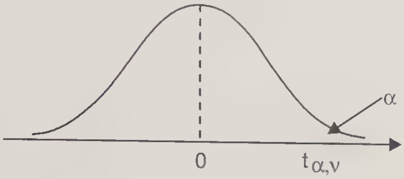


α	**Area in the Right Tail under the chi-square Distribution Curve**									
ν	**.995**	**.990**	**.975**	**.950**	**.900**	**.100**	**.050**	**.025**	**.010**	**.005**
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	5.844	7.633	8.907	10.117	11.651	27.204	20.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Appendix–G

The *t*-Distribution Table

Percentage Points $t_{\alpha,\nu}$ of the *t*-Distribution



ν \ α	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.638	2.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	.256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

ν = degrees of freedom.

Appendix–G (Continued)

The *t*-Distribution Table

df	Area in the Right Tail under the *t*-distribution Curve					
	0.1	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	3.078	6.314	12.706	31.821	63.657	318.309
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356
34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340
36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313
40	1.303	1.684	2.021	2.423	2.704	3.307

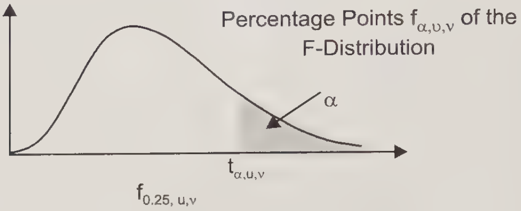
Appendix–G (Continued)

The *t*-Distribution Table

df	Area in the Right Tail under the *t*-distribution Curve					
	0.1	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
41	1.303	1.683	2.020	2.421	2.701	3.301
42	1.302	1.682	2.018	2.418	2.698	3.296
43	1.302	1.681	2.017	2.416	2.695	3.291
44	1.301	1.680	2.015	2.414	2.692	3.286
45	1.301	1.679	2.014	2.412	2.690	3.281
46	1.300	1.679	2.013	2.410	2.687	3.277
47	1.300	1.678	2.012	2.408	2.685	3.273
48	1.299	1.677	2.011	2.407	2.682	3.269
49	1.299	1.677	2.010	2.405	2.680	3.265
50	1.299	1.676	2.009	2.403	2.678	3.261
51	1.298	1.675	2.008	2.402	2.676	3.258
52	1.298	1.675	2.007	2.400	2.674	3.255
53	1.298	1.674	2.006	2.399	2.672	3.251
54	1.297	1.674	2.005	2.397	2.670	3.248
55	1.297	1.673	2.004	2.396	2.668	3.245
56	1.297	1.673	2.003	2.395	2.667	3.242
57	1.297	1.672	2.002	2.394	2.665	3.239
58	1.296	1.672	2.002	2.392	2.663	3.237
59	1.296	1.671	2.001	2.391	2.662	3.234
60	1.296	1.671	2.000	2.390	2.660	3.232
61	1.296	1.670	2.000	2.389	2.659	3.229
62	1.295	1.670	1.999	2.388	2.657	3.227
63	1.295	1.669	1.998	2.387	2.656	3.225
64	1.295	1.669	1.998	2.386	2.655	3.223
65	1.295	1.669	1.997	2.385	2.654	3.220
66	1.295	1.668	1.997	2.384	2.652	3.218
67	1.294	1.668	1.996	2.383	2.651	3.216
68	1.294	1.668	1.995	2.382	2.650	3.214
69	1.294	1.667	1.995	2.382	2.649	3.213
70	1.294	1.667	1.994	2.381	2.648	3.211
71	1.294	1.667	1.994	2.380	2.647	3.209
72	1.293	1.666	1.993	2.379	2.646	3.207
73	1.293	1.666	1.993	2.379	2.645	3.206
74	1.293	1.666	1.993	2.378	2.644	3.204
75	1.293	1.665	1.992	2.377	2.643	3.202
infinite	1.282	1.645	1.960	2.326	2.576	3.090

Appendix–H

The *F*-Distribution Table


Percentage Points $f_{\alpha,\upsilon,\nu}$ of the F-Distribution
α
$t_{\alpha,u,\nu}$
$f_{0.25,\ u,\nu}$

	ν \ α	Degrees of freedom for the numerator(u)																		
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**24**	**30**	**40**	**60**	**120**	**∞**
Degrees of freedom for the denominator (ν)	1	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.41	9.49	9.58	9.63	9.67	9.71	9.76	9.80	9.85
	2	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.41	3.43	3.43	3.44	3.45	3.46	3.47	3.48
	3	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47
	4	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08
	5	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.88	1.88	1.88	1.88	1.87	1.87	1.87
	6	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.76	1.76	1.75	1.75	1.75	1.74	1.74	1.74
	7	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.70	1.69	1.68	1.68	1.67	1.67	1.66	1.66	1.65	1.65	1.65
	8	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.62	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
	9	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.54	1.53	1.53
	10	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.54	1.53	1.52	1.52	1.51	1.51	1.50	1.49	1.48
	11	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47	1.46	1.45
	12	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.45	1.44	1.43	1.42
	13	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
	14	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.45	1.44	1.43	1.42	1.41	1.41	1.40	1.39	1.38
	15	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.43	1.41	1.41	1.40	1.39	1.38	1.37	1.36
	16	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
	17	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33
	18	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.34	1.33	1.32
	19	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.33	1.32	1.30
	20	1.40	1.49	1.48	1.47	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31	1.29
	21	1.40	1.48	1.48	1.46	1.44	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28
	22	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.37	1.36	1.34	1.33	1.32	1.31	1.30	1.29	1.28
	23	1.39	1.47	1.47	1.45	1.43	1.42	1.41	1.40	1.39	1.38	1.37	1.35	1.34	1.33	1.32	1.31	1.30	1.28	1.27
	24	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.36	1.35	1.33	1.32	1.31	1.30	1.29	1.28	1.26
	25	1.39	1.47	1.46	1.44	1.42	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.32	1.31	1.29	1.28	1.27	1.25
	26	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.35	1.34	1.32	1.31	1.30	1.29	1.28	1.26	1.25
	27	1.38	1.46	1.45	1.43	1.42	1.40	1.39	1.38	1.37	1.36	1.35	1.33	1.32	1.31	1.30	1.28	1.27	1.26	1.24
	28	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.34	1.33	1.31	1.30	1.29	1.28	1.27	1.25	1.24
	29	1.38	1.45	1.45	1.43	1.41	1.40	1.38	1.37	1.36	1.35	1.34	1.32	1.31	1.30	1.29	1.27	1.26	1.25	1.23
	30	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.34	1.32	1.30	1.29	1.28	1.27	1.26	1.24	1.23
	40	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.31	1.30	1.28	1.26	1.25	1.24	1.22	1.21	1.19
	60	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.27	1.25	1.24	1.22	1.21	1.19	1.17	1.15
	120	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.26	1.24	1.22	1.21	1.19	1.18	1.16	1.13	1.10
	∞	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.22	1.19	1.18	1.16	1.14	1.12	1.08	1.00

Appendix–H

The *F*-Distribution Table

Percentage Points of the *F*-Distribution (continued) –$f_{0.10,\,u,v}$
ν = Degrees of freedom for the denominator.

α	Degrees of freedom for the numerator(u)																		
ν	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83	1.78	1.73	1.68	1.65	1.62	1.58	1.55	1.51	1.47
30	2.88	2.49	2.28	2.14	2.03	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.64	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.51	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.45	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.55	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

Appendix–H

The *F*-Distribution Table

Percentage Points of the *F*-Distribution (continued) –$f_{0.05,\, u, v}$
ν = Degrees of freedom for the denominator.

ν \ α	Degrees of freedom for the numerator (u)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Appendix–H

The *F*-Distribution Table

Percentage Points of the *F*-Distribution (continued) $-f_{0.025,\,u,v}$
ν = Degrees of freedom for the denominator.

α	Degrees of freedom for the numerator (u)																		
ν	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

Appendix–H

The F-Distribution Table

Percentage Points of the F-Distribution (continued) $-f_{0.01,\,u,v}$
ν = Degrees of freedom for the denominator.

α / ν	Degrees of freedom for the numerator (u)																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.82	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.46
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.36	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.59
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.35	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Appendix–I

Cramer's Rule

Cramer's Rule: Cramer's rule can be used to solve a system of simultaneous linear algebraic equations. Consider a general system of n linear equations in n unknowns:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \vdots \qquad\quad \vdots \qquad\quad & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned} \tag{I.1}$$

and define the determinant of such a system. We begin by defining the determinant of a 3 × 3 system:

$$C = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \tag{I.2}$$

The general definition of the determinant of the $n \times n$ system of Equation (I.1) is simply an extension of the procedure (I.2).

$$C = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} = a_{11}A_{11} - a_{12}A_{12} + \cdots + (-1)^{n+1}a_{1n}A_{1n} \tag{I.3}$$

where A_{1j} is the $(n - 1) \times (n - 1)$ determinant obtained by crossing out the first row and j^{th} column of the original $n \times n$ determinant. Hence an $n \times 1$ determinant can be obtained by calculating $n(n - 1)(n - 1)$ determinants.

We shall conclude this appendix by introducing a method for obtaining solutions of the system (I.1). We define the determinants

$$C_1 = \begin{vmatrix} b_1 & a_{12} & \cdots & a_{1n} \\ b_2 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ b_n & a_{n2} & \cdots & a_{nn} \end{vmatrix},$$

$$C_2 = \begin{vmatrix} a_{11} & b_1 & a_{13} & \cdots & a_{1n} \\ a_{21} & b_2 & a_{23} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1} & b_n & a_{n3} & \cdots & a_{nn} \end{vmatrix}, \ldots,$$

$$C_k = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1,k-1} & b_1 & a_{1,k+1} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2,k-1} & b_2 & a_{2,k+1} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{n,k-1} & b_k & a_{n,k+1} & \cdots & a_{nn} \end{vmatrix}, \ldots,$$

$$C_n = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1,n-1} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2,n-1} & b_2 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & b_n \end{vmatrix}, \tag{I.4}$$

obtained by replacing the column of C by the column

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Then we have the following theorem, known as *Cramer's rule.*

Cramer's rule: Let C and C_k, $k = 1, 2, \ldots, n$, be given as in (I.3). If $C \neq 0$, then the unique solution to the system (I.1) is given by the values

$$x_1 = \frac{C_1}{C},\ x_2 = \frac{C_2}{C},\ \cdots,\ x_n = \frac{C_n}{C} \tag{I.5}$$

Example EI.1

Obtain the solution of the following simultaneous linear equations by Cramer's rule.

(a) $$\begin{bmatrix} 1 & 3 \\ 4 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 12 \end{bmatrix}$$

(b) $$\begin{bmatrix} 1 & -3 & 2 \\ 3 & 4 & 1 \\ -4 & 2 & -9 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ 5 \\ 2 \end{bmatrix}$$

SOLUTION:

(a) $$x_i = \frac{|C_i|}{|C|}$$

$$x_1 = \frac{\begin{vmatrix} 5 & 3 \\ 12 & -1 \end{vmatrix}}{\begin{vmatrix} 1 & 3 \\ 4 & -1 \end{vmatrix}} = \frac{-5-36}{-13} = 3.15$$

$$x_2 = \frac{\begin{vmatrix} 1 & 5 \\ 4 & 12 \end{vmatrix}}{\begin{vmatrix} 1 & 3 \\ 4 & -1 \end{vmatrix}} = \frac{12-20}{-13} = 0.62$$

(b)

$$x_i = \frac{|C_i|}{|C|}$$

$$x_1 = \frac{\begin{vmatrix} 8 & -3 & 2 \\ 5 & 4 & 1 \\ 2 & 2 & -9 \end{vmatrix}}{\begin{vmatrix} 1 & -3 & 2 \\ 3 & 4 & 1 \\ -4 & 2 & -9 \end{vmatrix}} = \frac{-441}{-63} = 7$$

$$x_2 = \frac{\begin{vmatrix} 1 & 8 & 2 \\ 3 & 5 & 1 \\ -4 & 2 & -9 \end{vmatrix}}{\begin{vmatrix} 1 & -3 & 2 \\ 3 & 4 & 1 \\ -4 & 2 & -9 \end{vmatrix}} = \frac{189}{-63} = -3$$

$$x_3 = \frac{\begin{vmatrix} 1 & -3 & 8 \\ 3 & 4 & 5 \\ -4 & 2 & 2 \end{vmatrix}}{\begin{vmatrix} 1 & -3 & 2 \\ 3 & 4 & 1 \\ -4 & 2 & -9 \end{vmatrix}} = \frac{252}{-63} = -4$$

Example EI.2

Consider the system

$$\begin{aligned} 2x_1 + 4x_2 - x_3 &= -5 \\ -4x_1 + 3x_2 + 5x_3 &= 14 \\ 6x_1 + 3x_2 - 2x_3 &= 5 \end{aligned}$$

We have

$$C = \begin{vmatrix} 2 & 4 & -1 \\ -4 & 3 & 5 \\ 6 & -3 & -2 \end{vmatrix} = 112 \qquad C_1 = \begin{vmatrix} -5 & 4 & -1 \\ 14 & 3 & 5 \\ 5 & -3 & -2 \end{vmatrix} = 224$$

$$C_2 = \begin{vmatrix} 2 & -5 & -1 \\ -4 & 14 & 5 \\ 6 & 5 & -2 \end{vmatrix} = -112 \qquad C_3 = \begin{vmatrix} 2 & 4 & -5 \\ -4 & 3 & 14 \\ 6 & -3 & 5 \end{vmatrix} = 560$$

Therefore $x_1 = \dfrac{C_1}{C} = 2$, $\quad x_2 = \dfrac{C_2}{C} = -1$, $\quad x_3 = \dfrac{C_3}{C} = 5$

PROBLEMS

Solve the following system of linear equations using Cramer's rule.

PI.1(*a*)

$$\begin{aligned} x + 2y + z &= 0 \\ 3x + y - z &= 0 \\ x - y + 4z &= 3 \end{aligned}$$

(*b*)

$$\begin{aligned} 2x + y - 3z &= 11 \\ 4x - 2y + 3z &= 8 \\ -2x + 2y - z &= -6 \end{aligned}$$

(*c*)

$$\begin{aligned} x + 3y - z &= -10 \\ -x + 4y + 2z &= -4 \\ 2x - 2y + 5z &= 35 \end{aligned}$$

(*d*)

$$\begin{aligned} y - 3z &= -5 \\ 2x + 3y - z &= 7 \\ 4x + 5y - 2z &= 10 \end{aligned}$$

(*e*)

$$\begin{bmatrix} 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 2 \\ 4 & 4 & 1 & 1 \\ 2 & 1 & 2 & -2 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 2 \\ 0 \\ 11 \\ 2 \end{Bmatrix}$$

(*f*)

$$\begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{Bmatrix} = \begin{Bmatrix} 32 \\ 23 \\ 33 \\ 31 \end{Bmatrix}$$

ANSWERS

(*a*) $x = 0.3334, y = -0.4444, z = 0.5555$

(*b*) $x = 3, y = -1, z = -2$

(*c*) $x = 2, y = -3, z = 5$

(*d*) $x = -1, y = 4, z = 3$

(*e*) $x_1 = 1, x_2 = 2, x_3 = -1, x_4 = 0$

(*f*) $x_1 = 1, x_2 = 1, x_3 = 1, x_4 = 1$

Bibliography

There are several outstanding text and reference books on *Probability and Statistics* that merit consultation for those readers who wish to pursue these topics further. The following list is a representative sample of the many excellent books.

Barr, D.R., and Zehna, P.W., *Probability: Modeling Uncertainty*, Addison Wesley, Reading, MA, 1983.

Bowker, A.H., and Lieberman, G.J., *Engineering Statistics*, 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1972.

Caella, G., and Berger, R.L., *Statistical Inference*, Brooks/Cole, Pacific Grove, CA, 2002.

Chambers, J., Cleveland, W., Kleiner, B., and Toukai, P., *Graphical Methods for Data Analysis*, Wadsworth and Brooks/Cole, Pacific Grove, CA, 1983.

Conover, W.J., *Practical Nonparametric Statistics*, 2nd ed., Wiley, New York, 1980.

DeGroot, M.H., and Schervish, M.J., *Probability and Statistics*, 3rd ed., Addison Wesley, Reading, MA, 1986.

Derman, C., Gleser, L. and Olkin, I., *Probability Models and Applications*, Macmillan, New York, 1980.

Devore, J.L., *Probability and Statistics for Engineering and the Sciences*, 4th ed., Wadsworth and Brooks/Cole, Pacific Grove, CA, 1995.

Dukkipati, R.V., *Probability and Statistics through Solved Problems*, New Age International Publishers (P) Ltd., New Delhi, India, 2009.

Eisen, M., *Elementary Combinatorial Analysis*, Gordon and Breach, Science Publishers, Inc., New York, 1970.

Freedman, D., Pisani, R., Purves, R. and Adhikari, A., *Statistics*, 2nd ed., Norton, New York, 1991.

Guttman, I., and Wilks, S.S., *Introductory Engineering Statistics*, Wiley, New York, 1965.

Hastings, N.A.J., and Peacock, J.B., *Statistical Distributions*, Butterworth & Co. Ltd., London, UK, 1975.

Hines, W.W., and Montgomery, D.C., *Probability and Statistics in Engineering and Management Science*, 3rd ed., Wiley, New York, 1990.

Hoel, P.G., *Introduction to Mathematical Statistics*, 5th ed., Wiley, New York, 1984.

Hoel, P.G., Port, S.C. and Stone, C.J., *Introduction to Probability Theory*, Houghton Mifflin, Boston, 1971.

Hogg, R.V., and Tannis, E.A., *Probability and Statistical Inference*, Prentice-Hall, Upper Saddle River, NJ, 1997.

Iman, R.L., *A Data-Based Approach to Statistics*, Duxbury Press, Belmont, CA, 1994.

Johnson, N.L., and Kotz, S., *Discrete Distributions*, Houghton-Mifflin Company, Boston, MA, 1969.

Johnson, R.A., Miller and Freund's *Probability and Statistics for Engineers*, 6th ed., Prentice-Hall of India Private Limited, New Delhi, India, 2000.

Kennedy, J.B. and Neville, A.M., *Basic Statistical Methods for Engineers and Scientist*s, 2nd ed., Harper and Row, New York, 1976.

Larsen, R., and Marx, M., *Introduction to Mathematical Statistics*, 3rd ed., Prentice-Hall, Englewood Cliffs, New Jersey, 2001.

Larson, H.J., *Introduction to Probability Theory and Statistical Inference*, 3rd ed., Wiley, New York, 1982.

Lindgren, B.W., *Statistical Theory,* Macmillan, New York, 1976.

Mann, P.S., *Introductory Statistics*, 3rd ed., Wiley, 1998.

Mason, R.L., Gunst, R.F., and Hess, J.L., *Statistical Design and Analysis of Experiments*, Wiley, New York, 1989.

Microsoft Excel User's Guide Version 5.0, *Microsoft,* Redmond, Washington, 1994.

Miller, I., and Miller, M., John E. Freund's *Mathematical Statistics with Applications*, 7th ed., Prentice-Hall of India Private Limited, New Delhi, India, 2007.

Minitab Reference Manual, Release 11, *Minitab,* State College, PA, 1996.

Moore, D.S., and MaCabe, G.P., *Introduction to the Practice of Statistics*, 3rd ed., W.H. Freeman, New York, 1999.

Mosteller, F., Rourke, R. and Thomas, G., *Probability with Statistical Applications*, 2nd ed., Addison-Wesley, Reading, MA, 1970.

Navidi, W.C., *Statistics for Engineers and Scientists*, 2nd ed., McGraw-Hill, New York, 2008.

Riordon, J., *An Introduction to Combinatorial Analysis*, Wiley, New York, 1958.

Rohagti, V.K., *An Introduction to Probability Theory and Mathematical Statistics,* Wiley, New York, 1976.

Ross, S., *A First Course in Probability*, 3rd ed., Macmillan, New York, 1985.

Ross, S., *Introduction to Probability and Statistics for Engineers and Scientists*, Wiley, New York, 1987.

Ryan, B.F., Joiner, B.L., and Ryan, T. A., *Minitab Handbook*, 2nd ed., PWS-Kent, Boston, 1985.

Stephens, L.J., *Engineering Statistics Demystified*, 1st ed., McGraw-Hill, New York, 2007.

Susan Milton, J., and Arnold, J.C., *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computer Science*, 4th ed., McGraw-Hill, New York, 2003.

Tanur, J., et al. (eds.), *Statistics: A Guide to the Unknown*, 3rd ed., Wadsworth and Brooks/Cole, Pacific Grove, CA, 1989.

Triola, M.F., *Elementary Statistics*, 3rd ed., Pearson Education, Inc, Boston, MA, 2006.

Tufte, E.R., *The Visual Display of Quantitative Information,* Graphics Press, Cheshire, Connecticut, 1983.

Tukey, J., *Exploratory Data Analysis*, Addison-Wesley, Reading, MA, 1977.

Walpole, R.E., Myers, R.H., and Myers, S.L., *Probability and Statistics for Engineers and Scientists*, 6th ed., Prentice-Hall, Upper Saddle River, New Jersey, 1998.

Weiss, N.A., *Introductory Statistics,* 8th ed., Pearson Education, Inc., New York, 2008.

Glossary of Technical Terms

Terminology used frequently in the field of Probability and Statistics is compiled here from various sources.

***t*-Distribution:** A continuous distribution with a specific type of bell-shaped curve with its mean equal to 0 and standard deviation equal to $\sqrt{df/(df-2)}$.

***z*-Value or *z*-Score:** The units of the standard normal distribution that are denoted by *z*.

β-Beta: The probability of not rejecting a null hypothesis when it actually is false. (The probability of committing a Type II error).

σ (Sigma): The significance level of a test of hypothesis that denotes the probability of rejecting a null hypothesis when it actually is true. (The probability of committing a Type I error).

α-Alpha: The probability that the statistical test will incorrectly reject the null hypothesis.

μ_d: The mean of the paired differences for the population.

σ_d: The standard deviation of the paired differences for the population.

d: The difference between two matched values in two samples collected from the same source. It is called the paired difference.

$\bar{d}$: The mean of the paired differences for a sample.

Absolute Dispersion: The actual variation or dispersion is called the absolute dispersion.

Absolute Error: The difference, without sign, between the actual value and a statistically determined model value.

all those value that would not be rejected with p =.05.

Alpha: The probability that the statistical test will incorrectly reject the null hypothesis.

Alternative Hypothesis: A claim about a population parameter that will be true if the null hypothesis is false.

Analysis of Variance (ANOVA): A statistical technique used to test whether the means of three or more populations are equal.

ANOVA: See analysis of variance.

Array: An array is an arrangement of raw numerical data in ascending or descending order of magnitude.

Average: Another name for the mean. The sum of all the data values divided by the number of data values.

Bar Graph: A graph made of bars whose heights represent the frequencies of respective categories.

Bernoulli Trial: Consider a process that consists of a sequence of n independent trials, where the outcome of each trial is a success or a failure. These are called Bernoulli trials.

Beta: The probability of making a *Type II error* when making a statistical hypothesis test. One minus this probability is power, the probability of correctly rejecting the null hypothesis.

Biased Estimator: See unbiased estimator.

Bimodal Distribution: A distribution that has two modes.

Binomial Experiment: An experiment that contains n identical trials such that each of these n trials has only two possible outcomes, the probabilities of these two outcomes remain constant for each trial, and the trials are independent.

Binomial Parameters: The total trails n and the probability of success p for the binomial probability distribution.

Binomial Probability Distribution: The probability distribution that gives the probability of x successes in n trials when the probability of success is p for each trial of a binomial experiment.

Box Plot: The box represents the middle 50% of the data values; that is, the data values from the first to third quartiles. The line or dot in the middle of the box denotes the median. Whiskers – lines extending from the box – extend either to the limits of the data or to a maximum length equal to 1.5 times the length of the box. In the latter case, points beyond the lines indicate outliers.

Box-and-Whisker Plot: A plot that shows the center, spread and skewness of a data set by drawing a box and two whiskers using the median, the first quartile, the third quartile, and the smallest and the largest values in the data set between the lower and the upper inner fences.

Categorical Variable: The values of a categorical variable are discrete categories. Usually the categories cannot be ordered from *highest* to *lowest* in any meaningful way. Rather, the categories just define different groups. Gender and religions are examples of categorical variables. Compare *continuous variable*. See also *variable*.

Census: A survey conducted by including every element of the population.

Central Limit Theorem: The theorem from which it is inferred that for a large sample size ($n \geq 30$), the shape of the sampling distribution of $\bar{x}$ is approximately normal. Also, by the same theorem, the shape of the sampling distribution of $\hat{p}$ is approximately normal for a sample for which $np > 5$ and $nq > 5$.

Chebyshev's Theorem: For any number k greater than 1, at least $(1 - 1/k^2)$ of the values for any distribution lie within k standard deviations of the mean.

chi-square Distribution: A distribution, with degrees of freedom as the only parameter, that is skewed to the right for small *df* and looks like a normal curve for large *df*.

chi-square Test: A statistical test for determining whether two categorical variables are related.

Class Boundary: The midpoint of the upper limit of one class and the lower limit of the next class.

Class Frequency: The number of values in a data set that belongs to a certain class.

Class Length: (See class size)

Class Midpoint or Mark: The class midpoint or mark is obtained by dividing the sum of the lower and upper limits (or boundaries) of a class by 2.

Class Size: The size or width of a class interval is the difference between the lower and upper class boundaries. It is also referred to as the class width, class size, or class length.

Class: An interval that includes all the values in a (quantitative) data set that fall within two numbers, the lower and upper limits of the class.

Classical Probability Rule: The method of assigning probabilities to outcomes or events of an experiment with equally likely outcomes.

Cluster Sampling: The sampling technique under which the population is divided into clusters and a sample is chosen from one or a few clusters.

Cluster: A subgroup (usually geographical) of the population that is representative of the population.

Coefficient of Determination: A measure that gives the proportion (or percentage) of the total variation in a dependent variable that is explained by a given independent variable.

Coefficient of Variation: A measure of relative variability that expresses standard deviation as percentage of the mean.

Combinations: The number of ways x elements can be selected from n elements.

Complementary Events: Two events that are taken together include all the outcomes for an experiment but do not contain any common outcomes.

Compound Event: An event that contains more than one outcome of an experiment. It is also called a *composite event*.

Conditional Probability: The probability of an event subject to the condition that another event has already occurred.

Confidence Interval: All the values for a parameter that would not be rejected if they were used as the null hypothesis in a statistical test of that parameter. A 95% confidence interval includes.

Confidence Level: Confidence level, denoted by $(1 - \alpha)$ 100%, states how much confidence we have that a confidence interval contains the true population parameter. This may be thought of as the probability that a given assertion is true. More precisely a statement is true at C confidence level if it is true for fraction C of all notional populations that give the same sampling result as the one actually observed.

Confidence Limit: Stated at C confidence level, the C confidence limits are the endpoints of the C confidence interval.

Consistent Estimator: A sample statistic with a standard deviation that decreases as the sample size increases.

Contingency Coefficient: Quantifies the degree of relationship between two categorical variables.

Continuity Correction Factor: Addition of .5 and/or subtraction of .5 from the value(s) of x when the normal distribution is used as an approximation to the binomial distribution, where x is the number of successes in n trials.

Continuous Data: Data, which can be described by a continuous variable, are called continuous data.

Continuous Random Variable: A random variable that can assume any value in one or more intervals. A random variable whose values are countable.

Continuous Variable: (see variable) A variable, which can theoretically assume any value between two given values, is called a continuous variable, otherwise it is called a discrete variable.

Control Group: The group on which no condition is imposed.

Controllable Factor: A design variable that is considered to influence the response and is included in the experiment. The experimenters can control its level.

Convenient Sample: A sample in which the most accessible members of the population are selected.

Covariance: Assesses the degree to which two variables co-vary or vary together. If the two variables are independent then the covariance will equal zero. It is computed as the mean of the products of the mean deviations for each variable.

Critical Value or Critical Point: One or two values that divide the whole region under the sampling distribution of *s* sample statistic into rejection and non-rejection regions.

Cross-Section Data: Data collected on different elements at the same point in time or for the same period of time.

Cumulative Frequency: The cumulative frequency of a class is the frequency of the class divided by the total frequency of all classes and is generally expressed as a percentage.

Cumulative Frequency Distribution: The total frequency of all values less than the upper class boundary of a given class interval is called the cumulative frequency. A table representing such cumulative frequencies is called a cumulative frequency distribution, cumulative frequency table or simply a cumulative distribution.

Cumulative Percentage: The cumulative relative frequency multiplied by 100.

Cumulative Probability: The sum of the probabilities from the minimum possible value up to a certain value.

Cumulative Probability Distribution: By cumulating probabilities we obtain cumulative probability distribution, which are analogous to cumulative relative frequency distributions. The function associated with this distribution is often called a distribution function.

Cumulative Relative Frequency: The cumulative frequency of a class divided by the total number of observations.

Curve Fitting: The general problem of finding equations of approximating curves, which fit gives sets of data, is called curve fitting.

Data or Data Set: Collection of observations or measurements on a variable.

Degrees of Freedom (*df*): The number of observations that can be chosen freely. For the estimation of μ using the *t*-distribution, the degrees of freedom are $n - 1$.

Degrees of Freedom for a Simple Linear Regression Model: Sample size minus 2, that is, $n - 2$.

Dependent Events: Two events for which the occurrence of one changes the probability of the occurrence of the other.

Dependent Variable: The variable to be predicted or explained.

Descriptive Statistics: Collection of methods that are used for organising, displaying and describing data using tables, graphs and summary measures.

Designed Experiment: A study in which the experiment controls the assignment of elements to different treatment groups.

Deterministic Model: A model in which the independent variable determines the dependent variable exactly. Such a model gives an exact relationship between two variables.

Discrete Data: Data, which can be described by a discrete variable, are called discrete data.

Discrete Probability Distribution: Of a variation x can assume a discrete set of values with respective probabilities where we can say that a discrete probability distribution for x has been defined.

Discrete Random Variable: A random variable whose values are countable.

Discrete Variable: A variable that can have only certain discrete values. The same as a categorical variable.

Dispersion: The degree to which numerical data tend to spread apart an average value is called the variation or dispersion of the data. The different measure of valuation or dispersion available is the range, mean deviation, semi-interquartile range, and the standard deviation.

Distribution Function: A function giving, for every value x, the probability that a random variable X is lower than equal to x.

Dot Plot: A simple graph in which each observation is represented by a dot at its score.

Efficient Estimator: If the sampling distributors of the two statistics have the same mean (or expectation) the statistic with the smaller variance is called an efficient estimator of the mean while the other statistic is called an inefficient estimator.

Element or Member: A specific subject or object included in a sample or population.

Empirical Rule: For a specific bell-shaped distribution, about 68% of the observations fall in the interval $(\mu - \sigma)$ to $(\mu + \sigma)$, about 95% fall in the interval $(\mu - 2\sigma)$ to $(\mu + 2\sigma)$, and about 99.7% fall in the interval $(\mu - 3\sigma)$ to $(\mu + 3\sigma)$.

Equally Likely Outcomes: Two (or more) outcomes or events that have the same probability of occurrence.

Error: The amount of variation in the response caused by factors other than controllable factors included in the experiment.

Estimate: The value of a sample statistic that is used to find the corresponding population parameter.

Estimated or Predicted Value of *y*: The value of the dependent variable, denoted by $\hat{y}$, that is calculated for a given value of x using the estimated regression model.

Estimation: Estimation is the process of analysing a sample result in order to predict the corresponding value of the population parameter. The sample statistic that is used to estimate a population parameter.

Estimator: The sample statistic that is used to estimate a population parameter.

Event: A collection of one or more outcomes of an experiment.

Expected Frequencies: The frequencies for different categories of a multinomial experiment or for different cells of a contingency table that are expected to occur when a given null hypothesis is true.

Experiment: A process with well-defined outcomes that, when performed, results in one and only one of the outcomes per repetition.

***F*-Distribution:** A probability distribution for ratios of variances. Most commonly used in *analysis of variance* to compare the ratio of the variance of the means from a number of groups were the same.

Factorial *n*: Factorial n, denoted by $n!$ is simply defined as $n! = n\ (n-1)\ (n-2)\cdots 1$.

Factorial: Denoted by the symbol "!"

First Quartile: The value in a ranked data set such that about 25% of the measurements are smaller than this value and about 75% are larger. It is the median of the values that are smaller than the median of the whole data set.

Frequency Distribution: A tabular arrangement of data by classes put together with the corresponding class frequencies is called a frequency distribution or frequency table.

Frequency Function: (See probability function).

Frequency Polygon: A frequency polygon is a line graph made by connecting the mid points of the bars in a histogram by straight lines.

Frequency Table: (See frequency distribution)

Frequency: The number of times a certain data value occurs in the set of observations.

Gaussian Distribution: Another name for the *normal distribution*.

Goodness-of-Fit Test: A test of the null hypothesis that the observed frequencies for an experiment follow a certain pattern or theoretical distribution.

Grouped Data: Data organised and summarised in a frequency-distributed manner are called grouped data.

Histogram: A graph in which classes are marked on the horizontal axis and frequencies, relative frequencies, or percentages are marked on the vertical axis. Bars that are drawn adjacent to each other represent the frequencies, relative frequencies, or percentages or various classes.

Hypothesis Test: A test of whether or not the null hypothesis should be rejected based on the data.

Hypothesis: A hypothesis is an assertion made about a population.

Impossible Event: An event that cannot occur.

Independent Events: Two events for which the occurrence of one does not change the probability of occurrence of the other.

Independent or Explanatory Variable: The variable included in a model to explain the variation in the dependent variable.

Independent Samples: Two samples drawn from two populations such that the selection of one does not affect the selection of the other.

Independent Variable: In an experiment, the levels of independent variables are controlled by the experimenter. Independent variables serve as inputs. Compare *dependent variable*.

Inefficient Estimator: See efficient estimator.

Inferential Statistics: Collection of methods that help make decisions about a population based on sample results.

Interaction: Two factors are said to have interaction with each other if the influence of one on the response function is dependent on the value of the other.

Intercept: In the following equation for a line such as a regression model, the parameter a is the intercept.

y-intercept: The point at which the regression line intersects the vertical axis on which the dependent variable is marked. It is the value of y when x is zero.

Interquartile Range (IQR): The difference between the third Quartile ($Q3$) and the first quartile ($Q1$). The IQR indicates the range of scores spanned n by the middle half of the data values. The box in a box plot represents the IQR.

Intersection of Events: Intersection of events is given by the outcomes that are common to two (or more) events.

Interval Estimate: An estimate of a population parameter given by two numbers between which the parameter may be considered to lie is called an interval estimate of the parameter.

Joint Probability: The probability that two (or more) events occur together.

Judgment Sample: A sample that includes the elements of the population selected based on the judgment and prior knowledge of an expert.

Kruskal-Wallis Test: A non-parametric statistical procedure for comparing medians across groups.

Kurtosis: Kurtosis is the degree of peaked ness of a distribution relative to a normal distribution.

Large Sample: The samples of size greater than 30 are called large samples.

Law of Large Numbers: If an experiment is repeated again and again, the probability of an event obtained from the relative frequency approaches the actual or theoretical probability.

Least-Squares Estimates of a and b: The values of a and b that are calculated by using the sample data.

Least-Squares Method: The method used to fit a regression line through a scatter diagram such that the error sum of squares is a minimum.

Least-Squares Regression Line: A regression line obtained by using the least squares method.

Left-Tailed Test: A test in which the rejection region lies in the left tail of the distribution curve.

Level of Significance: The maximum probability in testing a hypothesis is with which type I error is considered to be in risk is called the level of significance of the test.

Linear Correlation Coefficient: A measure of the strength of the linear relationship between two variables.

Linear Regression Model: A regression model that gives a straight-line relationship between two variables.

Lower Inner Fence: The value in a data set that is 1.5 × IQR below the first quartile.

Lower Outer Fence: The value in a data set that is 3.0 × IQR below the first quartile.

Lower Quartile: The lowest 25% of the scores are equal to or less than the lower quartile score. Often represented as Q1; equal to the 25th *percentile*.

Marginal Probability: The probability of one event or characteristic without consideration of any other event.

Maximum Error of Estimate: The quantity that is subtracted from and added to the value of a sample statistic to obtain a confidence interval for the corresponding population parameter.

Maximum: The largest data value in a set of observations.

Mean of a Discrete Random Variable: The mean of a discrete random variable x is the value that is expected to occur per repetition, on average, if an experiment is performed a large number of times. The mean of a discrete random variable is also called its *expected value*.

Mean of $\hat{p}$: The mean of the sampling distribution of $\bar{x}$, denoted $\mu_{\bar{x}}$, is equal to the population mean μ.

Mean Square Between Samples or MSB: A measure of the variation among means of samples taken from different populations.

Mean Square Within Samples or MSW: A measure of the variation within data of all samples taken from different populations.

Mean: The average of the data values, that is, the sum of all the data values divided by the number of data values. The mean makes the *sum of squared errors* as small as possible.

Measure of Dispersion: Measures that give the spread of a distribution. The range, variance and standard deviation are three such measures.

Measures of Central Tendency: Measures that describe the centre of a distribution. The mean, median and mode are three of the measures of central tendency.

Measures of Position: Measures that determine the position of a single value in relation to other values in a data set. Quartiles, percentiles, and percentile rank are examples of measures of position.

Median Absolute Deviation (MAD): The median of all the absolute errors or deviations. The MAD describes the spread of the data values away from the model value.

Median: The middle data value in a set of observation. To find the median, reorder the data from smallest to largest and find the middle observation; that is the median. If there is an even number of observations, then there will be two middle values; in that case, the average of those two middle values is the median. The median is also the 50th *percentile.*

Minimum: The lowest data value in a set of observations.

Mode: A value (or values) that occurs with highest frequency in a data set.

Model: Specifies a predicted or representative value of reach observation in a data set. Simple models are the mean or the median for a set of observations. More complex models are, for example, regression lines.

Multimodal Distribution: A distribution that has more than two modes.

Multinomial Experiment: An experiment with n trials for which (1) the trials are identical, (2) there are more than two possible outcomes per trial, (3) the trials are independent and (4) the probabilities of various outcomes remain constant for each trial.

Multiple Regression Model: A regression model that contains two or more independent variables.

Mutually Exclusive Events: Two or more events that do not contain any common outcome and, hence, cannot occur together.

Negative Relationship between Two Variables: The value of the slope in the regression line and the correlation coefficient between two variables are both negative.

Nominal Variable: Same as *categorical variable*. Levels of these variables are *noms* or *names*. For example, the names for the levels of the gender variable are *male* and *female.*

Non-linear (simple) Regression Model: A regression model that does not give a straight-line relationship between two variables.

Non-parametric Statistics: Statistical methods that make no assumption that the data are from a *normal distribution.*

Non-random Sample: A sample in which some members of the population may have no chance of being selected.

Non-sampling Errors: The errors that occur during the collection, recording, and tabulation of data.

Non-sampling or Systematic Error: The errors that occur in the collection, recording and tabulation of data.

Normal Distribution: The probability distribution that describes or approximates the distribution of many variables. Most observations in a normal distribution occur near the mean.

Normal Probability Distribution: The probability distribution of a continuous random variable that, when plotted, gives a specific bell-shaped curve. The parameters of the normal distribution are the mean μ and the standard deviation σ.

Null Hypothesis: The hypothesis that nothing is happening in the data that would not be expected by chance. Some null hypotheses are (a) the means of two groups are equal, (b) the slope in a simple regression is zero, (c) the correlation between two variables is zero, and (d) the classification of the data according to one nominal variable is independent of the classification of that data according to a second nominal variable.

Number of Degrees of Freedom: The number of degrees of freedom (n) of a statistic generally is defined as the number N of the independent observatory in the sample (sample size) minus the number k of population parameters estimated from the sample observations.

Observation or Measurement: The value of a variable for an element.

Observation Study: A study in which the assignment of elements to different treatment is voluntary and the researcher simply observes the results of the study.

Observation: A person, place, or thing for which a data value is available or is measured.

Observed Frequencies: The frequencies actually obtained from the performance of an experiment.

Observed Value of z or t: The value z or t calculated for a sample statistic such as the sample mean or the sample proportion.

Ogive: A curve drawn for a cumulative frequency distribution.

One-Way ANOVA: The analysis of variance technique that analyses one variable only.

One-Tailed Test: A test in which there is only one rejection region, either in the left tail or in the right tail of the distribution curve.

Ordinal Variable: A variable whose values indicate only rank and not the distance between values. For example, in a race, an ordinal variable would indicate who was first, second, third, etc., but would not give the time or distance between the racers.

Outcome: An outcome is one particular result of an experiment.

Outlier: An extreme or a typical value that is very different from other data values in the set of observations.

Outliers or Extreme Values: Values that is very small or very large relative to the majority of the values in a data set.

Paired or Matched Samples: Two samples drawn in such a way that they include the same elements and two data values are obtained from each element, one for each sample. Also called dependent samples.

Parameter: An unknown value in a model that is estimated from the data. Examples: mean, median, intercept, and slope.

Percentage: The percentage for a class or category is obtained by multiplying the relative frequency of that class or category by 100.

Percentile Rank: The percentile rank of a value gives the percentage of values in the data set that are smaller than this value.

Percentile: The P^{th} percentile data value is the score that is equal to or greater than $P\%$ of all the data values. For example, the score equal to the 25th percentile is equal to or greater than one quarter of all the data values (and is also the first *quartile*). The 50th percentile necessarily is the median. Ninety-nine values that divide a ranked data set into 100 equal parts.

Permutations: A permutation of n different objects taken r at a time is an arrangement of r out of the n objects with no special attention given to the order of arrangement. The number of combinations of n objects taken r at a time is denoted by np_r and is given by $np_r = \dfrac{n!}{(n-r)!}$

Pie Chart: A circle divided into portions that represent the relative frequencies or percentages of different categories or classes.

Point Estimate: An estimate of a population parameter given by a single number is called a point estimate of the parameter.

Poisson Parameter: The average occurrences, denoted by λ, during an interval for a Poisson probability distribution.

Poisson Probability Distribution: The probability distribution that gives the probability of x occurrences in an interval when the average occurrences in that interval are λ.

Polygon: A graph formed by joining the midpoints of the tops of successive bars in a histogram by straight lines.

Population Distribution: The probability distribution of the population data.

Population or Target Population: The collection of all elements whose characteristics are being studied.

Population Parameters for a Simple Regression Model: The values of A and B for the regression model $y = A + Bx + \in$ that are obtained by using population data.

Population Parameters: The estimation of unknown population quantities (such as population mean, variance, etc.) often called population parameters from a knowledge of corresponding sample quantities (sample mean variance, etc.), often called sample statistics.

Population Proportion p: The ratio of the number of elements in a population with a specific characteristic to the total number of elements in the population.

Population: A population consists of all the individuals or objects in a well-defined group about which information is desired to answer a question.

Positive Relationship between Two Variables: The value of the slope in the regression line and the correlation coefficient between two variables are both positive.

Power: The probability that a statistical test will reject the *null hypothesis* given that some specific alternative model is true. Equivalently, the probability that a *Type II error* will be avoided.

Prediction Interval: The confidence interval for a particular value of y for a given value of x.

Probabilistic or Statistical Model: A model in which the independent variable does not determine the dependent variable exactly.

Probability: The probability of an event occurring in an experiment equals the proportion of times the event occurs if the experiment is repeated a very large (infinite) number of times.

Probability Distribution of a Discrete Random Variable: A list of all the possible values that a discrete random variable can assume and their corresponding probabilities.

Probability Distribution: A distribution that determines the probability that a random variable takes any given value or belongs to a given set of values.

Qualitative or Categorical Variable: A variable that cannot assume numerical values but is classified into two or more categories.

Quantitative Data: Data generated by a quantitative variable.

Quartiles: Summary measures that divide all the scores into four quarters. One-fourth of the data values are equal to or less than the first or *lower quartile* ($Q1$), one half of the values are equal to or less than the second quartile ($Q2$, also known as the *median*), and three fourths of all the data values are equal to or less than the third or *upper quartile* ($Q3$).

Quota Sample: A sample selected in such a way that each group or subpopulation is represented in the sample in exactly the same proportion as in the target population.

Random Error Term ($\in$): The difference between the actual and predicted value of y.

Random Sample: A sample drawn in such a way that each element of the population has same chance of being included in the sample.

Random Sampling: Each member of a population, which has an equal chance of being included, is called random sampling.

Random Variable: A variable, denoted by x, whose value is determined by the outcome of a random experiment. Also called a *chance variable.*

Randomisation: The procedure under which elements are assigned to different (treatment and control) groups at random.

Range: The difference between the largest and smallest numbers is called the range of the data.

Ranks: Data values, which indicate only the order of the observations. See *ordinal variable.*

Raw Data: Data recorded in the sequence in which they are collected and before they are processed.

Relative Cumulative Frequency: The relative cumulative frequency or percentage cumulative frequency is the cumulative frequency divided by the total frequency.

Relative Dispersion: Relative dispersion is defined as the ratio of absolute dispersion and the average value.

Relative Frequency: The frequency of a class or category divided by the sum of all frequencies.

Representative Sample: A sample that contains the characteristics of the corresponding population.

Right-Tailed Test: A test in which the rejection region lies in the right rail of the distribution curve.

Rules of Decision: (See Tests of Significance).

Sample: A selection of observations from a larger population of all possible observations. Ideally, the sample is selected randomly from the population.

Sample Point: An outcome of an experiment.

Sample Proportion $\hat{p}$: The ratio of the number of elements in a sample with a specific characteristic to the total number of elements in that sample.

Sample Space: A sample space is the set of all possible outcomes of an experiment. These outcomes are also called the elements of the sample space.

Sample Statistics: See population parameters.

Sample Survey: A survey that includes elements of a sample.

Sampling Distribution: The probability distribution of the statistic one would get by sampling over and over again.

Sampling Distribution of $\bar{x}$: The probability distribution of all the values of $\bar{x}$ calculated from all possible samples of the same size selected from a population.

Sampling Distribution of $\hat{p}$: The probability distribution of all the values of $\hat{p}$ calculated from all possible samples of the same size selected from a population.

Sampling Error: The difference between the values of a sample statistic calculated from a random sample and the value of the corresponding population parameter. This type of error occurs due to chance.

Sampling Frame: The list of elements of the target population that is used to select a sample.

Sampling or Chance Error: The difference between the results obtained from a sample survey and the result that would be obtained from the census.

Scatter Diagram or Scattergram: A plot of the paired observations of x and y.

Scatter Plot, Scattergram: A two-variable plot in which each point represents the levels of an observation on each variable. Often the variable on the horizontal axis is to be used to predict the variable on the vertical axis. Scatter plots are useful for viewing *correlations* and for *simple regression.*

z-Score Formula: The difference between the observation and the mean, divided by the standard deviation.

S_d: The standard deviation of the paired differences for a sample.

Second Quartile: Middle or second of the three quartiles that divide a ranked data set into four equal parts. About 50% of the values in the data set are smaller and about 50% are larger than the second quartile. The second quartile is the same as the median.

Selection Error: The error that occurs because the sampling frame is not representative of the population.

Significance Level: The pre-specified probability of a Type I error below which the null hypothesis is considered to be so rare as to be implausible.

Significant: When the probability of obtaining a statistical value is below the significance level, the statistic is said to be *statistically significant.*

Simple Event: An event that contains one and only one outcome of an experiment. It is also called an *elementary event.*

Simple Linear Regression: A regression model with one dependent and one independent variable that assumes a straight-line relationship.

Simple Random Sample: A sample chosen in such a way that each element of the population has the same probability of being included in the sample.

Sixtieth Percentile: The sixtieth percentile is larger than 60% of the scores (and smaller than 40%).

Skewed: A skewed distribution is not symmetric. That is, it has different shapes on each side of the median. For example:

Skewed-to-the-Left Histogram: A histogram with a longer tail on the left side.

Skewed-to-the-Right Histogram: A histogram with a longer tail on the right side.

Skewness: Skewness is the degree of asymmetry or departure symmetry of a distribution. Of the frequency curve of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or 5 to have positive skewness. Of the reverse is true, it is called skewed to the left or to have negative skewness.

Slope: The coefficient of x in a regression model that gives the change in y for a change of one unit in x.

Small Samples: The samples of site lesser than 30 are called small samples.

SSB: The sum of squares between samples. Also called the sum of squares of the factor or treatment.

SSE (error sum of squares): The sum of the squared differences between the actual and predicted values of y. It is the portion of the SST that is not explained by the regression model.

SSR (regression sum of squares): The portion of the SST that is explained by the regression model.

SST: The total sum of squares given by the sum of SSB and SSW.

SST (total sum of squares): The sum of the squared differences between actual y value and $\bar{y}$.

SSW: The sum of squares within samples. Also called the sum of squares of errors.

Standard Deviation: A measure of spread that is given by the positive square root of the variance.

Standard Deviation of $\bar{x}$: The standard deviation of the sampling distribution of $\bar{x}$, denoted by $\sigma_{\bar{x}}$, is equal to $\sigma/\sqrt{n}$ when $n/N \leq .05$.

Standard Deviation of a Discrete Random Variable: A measure of spread for the probability distribution of a discrete random variable.

Standard Deviation of Errors: A measure of spread for the random errors.

Standard Deviation of $\hat{p}$: The standard deviation of the sampling distribution of $\hat{p}$, denoted by $\sigma_{\hat{p}}$, is equal to $\sigma/\sqrt{n}$ when $n/N \leq .05$.

Standard Error: The standard deviation of the sampling distribution for any given statistic.

Standard Error of Mean: The standard deviation of the sampling distribution of the mean.

Standard Normal Distribution: The normal distribution with $\mu = 0$ and $\sigma = 1$. The units of the standard normal distribution are denoted by z.

Statistic: A summary measure calculated for sample data.

Statistics: Statistics deals with scientific methods for collecting, organising, summarising, presenting and analysing data, as well as drawing valid conclusions and making reasonable decisions based on such analysis.

Statistical Hypothesis: Assumptions or guesses made which may or may not be true, are called statistical hypotheses and in general are statements about the probability distributions of the populations.

Stem-and-Leaf Display: A display of data in which each value is divided into two portions, a stem and a leaf.

Stratified Random Sampling: The sampling technique under which the population is divided into different strata and a sample is chosen from each stratum.

Stratum: A subgroup of the population whose members are identical with regard to the possession of a characteristic.

Subjective Probability: The probability assigned to an event based on the information and judgment of a person.

Sure Event: An event that is certain to occur.

Survey: Collecting data on the elements of a population or sample.

Symmetric: A symmetric distribution has the same shape on each side of the median. The normal distribution is an example of a symmetric distribution. Compare *skewed.*

Symmetric Histogram: A histogram that is identical on both sides of its central point.

Systematic Random Sampling: Sampling method used to choose a sample by selecting every kth unit from the list.

Target Population: The collection of all subjects of interest.

Target Value: A value that a product is expected to posses. Most often this value is different from what a single unit actually does exhibit. For a 9 volt transistor battery the target value is 9 volts.

Test of Homogeneity: A test of the null hypothesis that the proportions of elements that belong to different groups in two (or more) populations are similar.

Tests of Hypothesis: Procedures, which will enable to decide whether to accept or reject hypothesis or to determine whether observed samples differ significantly from expected results are called tests of hypothesis; tests of significance, or rules of decision.

Test of Independence: A test of the null hypothesis that attributes of a population is not related.

Test Statistic: The value of z or t calculated for a sample statistic such as the sample mean or the sample proportion.

Tests of Significance: (see tests of significance)

Third Quartile: Third of the three quartiles that divide a ranked data set into four equal parts. About 75% of the values in a data set are smaller than the value of the third quartile and about 25% are lager. It is the median of the values that are greater than the median of the whole data set.

Tree Diagram: A diagram in which each outcome of an experiment is represented by a branch of a tree.

Trimmed Mean: The $k\%$ trimmed mean is obtained by dropping $k\%$ of the smallest values and $k\%$ of the largest values from the given data and then calculating the mean of the remaining $(100 - 2k)\%$ of the values.

Two-Tailed Test: A two-tailed test considers whether the null hypothesis might be wrong in either direction; that is, it allows the possibility that the true parameter might be either above or below the value of the null hypothesis. Compare *one-tailed test.*

Type I Error: An error that occurs when a true null hypothesis is rejected.

Type II Error: An error that occurs when a false null hypothesis is not rejected.

Unbiased Estimator: Of the mean of the sampling distribution of a statistic equals to the corresponding population parameters, the statistic is called an unbiased estimator of the parameters, and otherwise it is called a biased estimator.

Uniform or Rectangular Histogram: A histogram with the same frequency for all classes.

Unimodal Distribution: A distribution that has only one mode.

Union of Two Events: All outcomes that belong either to one or to both events.

Upper Inner Fence: The value in a data set that is 1.5 × IQR above the third quartile.

Upper Outer Fence: The value in a data set that is 3.0 × IQR above the third quartile.

Upper Quartile: The 75th *percentile*. The observation that is greater than or equal to 75% of the observations and less than 25% of the observations. Sometimes designated by Q_3.

***p*-Value:** The smallest significance level at which a null hypothesis can be rejected.

Variable: A characteristic, trait, attribute, or measurement that can take on different values. A variable must vary, having at least one value for some observations and another value for other observations.

Variables, Factors, or Parameters: These words are used synonymously to indicate the controllable factors in an experiment. In the case of a plastic molding experiment, molding temperature, injection pressure, set time, etc., are factors.

Variance: The typical or average squared error. For a population, the sum of squared errors is divided by n the number of observations. More often, interest is in the sample estimate of the variance in which case the sum of squared errors is divided by $n - 1$.

Variation: —See dispersion—

Venn Diagram: A picture that represents a sample space or certain events.

***y*-intercept:** The point at which the regression line intersects the vertical axis on which the dependent variable is marked. It is the value of y when x is zero.

Glossary of Symbols

A	(1) y-intercept in the population regression model; (2) event A
a	y-intercept in the regression model estimated for sample data
$\overline{A}$	Complement of event A
α	(Greek letter alpha). Probability of Type I error or significance level
$(1-\alpha)$	Confidence level
B	Slope of the population regression line
b	Slope of the regression model estimated for sample data
β	(Greek letter beta). Probability of Type II error
C	Number of columns in a contingency table
χ^2	(χ is Greek letter chi). chi-square distribution
d	Difference between two paired values
$\overline{d}$	Mean of the paired differences for a sample
df	Number of degrees of freedom
E	(1) Maximum error of estimate; (2) Expected frequencies
E_i	Simple event i for an experiment
e	(1) Random error, $y - \bar{y}$, for a regression model; (2) $e = 2.71828$ in the Poisson probability distribution
$E(x)$	Expected value of a random variable x
ε	(Greek letter epsilon). Random error term in the population regression model
f	Frequency of a category or class
F	The F-distribution
H_0	Null hypothesis

H_1	Alternative hypothesis
IQR	Inter quartile range
k	(1) Number of different samples or treatments in a one-way ANOVA; (2) Number of categories in a multinomial experiment
λ	(Greek letter lambda). Mean of the Poisson probability distribution
μ	Midpoint of a class
μ	(Greek letter mu). Population mean
μ_b	Mean of the sampling distribution of b
μ_d	Mean of the paired differences for the population
$\mu_{\bar{d}}$	Mean of the sampling distribution of $\bar{d}$
$\mu_{\bar{x}}$	Mean of the sampling distribution of $\bar{x}$
$\mu_{\hat{p}}$	Mean of the sampling distribution of $\hat{p}$
$\mu_{y\|x}$	Mean of y for a specific value of x using population regression model
n	Sample size
$n!$	n factorial
$\binom{n}{x}$	Number of combinations of n elements selected x at a time
N	Population size
O	Observed frequencies for a category or cell
p	(1) Population proportion; (2) Probability of success for a binomial experiment
$\hat{p}$	Sample proportion
$\bar{p}$	Pooled sample proportion for two samples
P_k	kth percentile, where k is an integer from 1 to 99
$P(A)$	Probability that event A will occur
$P(A \text{ and } B)$	Joint probability of events A and B
$P(A \text{ or } B)$	Probability of the union of events A and B
$P(A \mid B)$	Probability that event A will occur given event B has occurred
q	$1 - p$, probability of failure for a binomial experiment
$\hat{q}$	$1 - \hat{p}$,where $\hat{p}$ is the sample proportion
$\bar{q}$	$1 - \bar{p}$, where $\bar{p}$ is the pooled sample proportion for two samples
Q_1, Q_2, Q_3	First, second, and third quartiles, respectively
R	Number of rows in a contingency table

r	Linear correlation coefficient for sample data
r^2	Coefficient of determination
ρ	(Greek letter rho). Linear correlation coefficient for population data
S	Sample space
s	Sample standard deviation
s^2	Sample variance
s_b	Estimator of σ_b
s_d	Standard deviation of the paired differences for a sample
$s_{\bar{d}}$	Estimator of $\sigma_{\bar{d}}$
s_e	Standard deviation of errors for the sample regression model
s_p	Pooled standard deviation
$s_{\hat{p}}$	Estimator of $\sigma_{\hat{p}}$
$s_{\hat{p}_1-\hat{p}_2}$	Estimator of $\sigma_{\hat{p}_1-\hat{p}_2}$
$s_{\bar{x}}$	Estimator of $\sigma_{\bar{x}}$
$s_{\bar{x}_1-\bar{x}_2}$	Estimator of $\sigma_{\bar{x}_1-\bar{x}_2}$
$s_{\hat{y}_m}$	Estimator of $\sigma_{\hat{y}_m}$
$s_{\hat{y}_p}$	Estimator of $\sigma_{\hat{y}_p}$
Σ	(Greek letter capital sigma). Summation notation
σ	(Greek letter lower-case sigma). Population standard deviation
σ^2	Population variance
σ_b	Standard deviation of the sampling distribution of b
σ_d	Standard deviation of the paired differences for the population
$\sigma_{\bar{d}}$	Standard deviation of the sampling distribution of $\bar{d}$
σ_ε	Standard deviation of errors for the population regression model
$\sigma_{\hat{p}}$	Standard deviation of the sampling distribution of $\hat{p}$
$\sigma_{\hat{p}_1-\hat{p}_2}$	Standard deviation of the sampling distribution of $\hat{p}_1 - \hat{p}_2$
$\sigma_{\bar{x}}$	Standard deviation of the sampling distribution of $\bar{x}$
$\sigma_{\bar{x}_1-\bar{x}_2}$	Standard deviation of the sampling distribution of $\bar{x}_1 - \bar{x}_2$
$\sigma_{\hat{y}_m}$	Standard deviation of $\hat{y}$ when estimating $\mu_{y\|x}$

$\sigma_{\hat{y}_p}$	Standard deviation of $\hat{y}$ when predicting y_p
t	The t-distribution
T_i	Sum of the values included in sample i in one-way ANOVA
x	(1) Variable; (2) Random variable; (3) Independent variable in a regression model
$\overline{x}$	Sample mean
y	(1) Variable; (2) Dependent variable in a regression model
$\hat{y}$	Estimated or predicted value of y using a regression model
z	Units of the standard normal distribution

Key Formulae

Notation: The following notation is used.

n = sample size

$\overline{x}$ = sample mean

s = sample std. dev.

Q_j = j^{th} quartile

N = population size

μ = population mean

σ = population std. dev.

d = paired difference

$\hat{p}$ = sample proportion

p = population proportion

O = observed frequency

E = expected frequency

Chapter 1: Numerical Descriptive Measures

- Relative frequency of a class = $f/\Sigma f$
- Percentage of a class = (relative frequency) × 100
- Class midpoint or mark = (upper limit + lower limit)/2
- Class width = upper boundary – lower boundary
- Cumulative relative frequency = $\dfrac{\text{cumulative frequency}}{\text{total observations in the data set}}$
- Cumulative percentage = (cumulative relative frequency) × 100
- Mean for ungrouped data: $\mu = \Sigma x/N$ and $\overline{x} = \Sigma x/n$
- Mean for grouped data: $\mu = \Sigma mf/N$ and $\overline{x} = \Sigma mf/n$

where μ is the midpoint and f is the frequency of a class

- Median for ungrouped data = value of the $\left(\frac{n+1}{2}\right)$th term in a ranked data set
- Range = largest value – smallest value
- Variance for ungrouped data

$$\sigma^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N} \quad \text{and} \quad s^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

where σ^2 is the population variance and s^2 is the sample variance

- Standard deviation for ungrouped data

$$\sigma = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{N}}{N}} \quad \text{and} \quad s = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$$

where σ and s are the population and sample standard deviations, respectively

- Variance for grouped data

$$\sigma^2 = \frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{N}}{N} \quad \text{and} \quad s = \frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{n}}{n}$$

- Standard deviation for grouped data

$$\sigma = \sqrt{\frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{N}}{N}} \quad \text{and} \quad s = \sqrt{\frac{\Sigma m^2 f - \frac{(\Sigma mf)^2}{n}}{n-1}}$$

- Sample standard deviation

$$s = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n-1}} \quad \text{or} \quad s = \sqrt{\frac{\Sigma x_i^2 - (\Sigma x_i)^2 / n}{n-1}}$$

- Population mean (mean of a variable: $\mu = \frac{\Sigma x_i}{N}$)
- Population standard deviation (standard deviation of a variable)

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}} \quad \text{or} \quad \sigma = \sqrt{\frac{\Sigma x_i^2}{N} - \mu^2}$$

- Standardised variable: $z = \frac{x - \mu}{\sigma}$
- Chebyshev's theorem

  For any number k greater than 1, at least $(1 - 1/k^2)$ of the values for any distribution lie within k standard deviations of the mean.

- Empirical rule

  For a specific bell-shaped distribution, about 68% of the observations fall in the interval $(\mu - \sigma)$ to $(\mu + \sigma)$, about 95% fall in the interval $(\mu - 2\sigma)$ to $(\mu + 2\sigma)$, and about 99.7% fall in the interval $(\mu - 3\sigma)$ to $(\mu + 3\sigma)$.
- Interquartile range: $\text{IQR} = Q_3 - Q_1$

  where Q_3 is the third quartile and Q_1 is the first quartile
- Lower limit $= Q_1 - 1.5\ \text{IQR}$
- Upper limit $= Q_3 + 1.5\ \text{IQR}$
- The k^{th} percentile

$$P_k = \text{value of the } \left(\frac{kn}{100}\right) \text{th term in a ranked data set}$$

- Percentile rank of $x_i = \dfrac{\text{number of values less than } x_i}{\text{total number of values in the data set}} \times 100$

Chapter 2: Probability

- Classical probability rule for a simple event

$$P(E_i) = \frac{1}{\text{total number of outcomes}}$$

- Classical probability rule for a compound event

$$P(A) = \frac{\text{number of outcomes in } A}{\text{total number of outcomes}}$$

- Relative frequency as an approximation of probability

$$P(A) = \frac{f}{n}$$

- Conditional probability of an event

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} \quad \text{and} \quad P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

- Condition for independence of events

$$P(A) = P(A|B) \quad \text{and/or} \quad P(B) = P(B \mid A)$$

- For complementary events: $P(A) + P(\bar{A}) = 1$
- Multiplication rule for dependent events

$$P(A \text{ and } B) = P(A)\, P(B|A)$$

- Multiplication rule for independent events

$$P(A \text{ and } B) = P(A)\, P(B)$$

- Joint probability of two mutually exclusive events

$$P(A \text{ and } B) = 0$$

- Addition rule for mutually nonexclusive events

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

- Addition rule for mutually exclusive events

$$P(A \text{ or } B) = P(A) + P(B)$$

- Special addition rule

$$P(A \text{ or } B \text{ or } C \text{ or } ...) = P(A) + P(B) + P(C) + \cdots \ (A, B, C, ... \text{ mutually exclusive})$$

- Complementation rule: $P(E) = 1 - P(\text{not } E)$
- General addition rule: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
- General multiplication rule: $P(A \text{ and } B) = P(A) \cdot P(B|A)$
- Special multiplication rule: $P(A \text{ and } B \text{ and } C \text{ and } \ldots) = P(A) \cdot P(B) \cdot P(C) \cdots$ $(A, B, C, \ldots$ independent)
- Rule of total probability

$$P(B) = \sum_{j=1}^{k} P(A_j) \cdot P(B \mid A_j)$$

(A_1, A_2, ..., A_k mutually exclusive and exhaustive)

- Bayes's rule

$$P(A_i \mid B) = \frac{P(A_i) \cdot P(B \mid A_i)}{\sum_{j=1}^{k} P(A_j) \cdot P(B \mid A_j)}$$

(A_1, A_2, ..., A_k mutually exclusive and exhaustive)

- Factorial: $k! = k(k-1) \cdots 2 \cdot 1$
- Permutations rule: $mP_r = \dfrac{m!}{(m-r)!}$
- Special permutations rule: $mP_m = m!$
- Combinations rule: $mC_r = \dfrac{m!}{r!(m-r)!}$
- Number of possible samples: $NC_n = \dfrac{N!}{n!(N-n)!}$

Chapter 3: Random Variables and Probability Distributions

- Mean of a discrete random variable X: $\mu = \Sigma x P(X = x)$
- Standard deviation of a discrete random variable X

$$\sigma = \sqrt{\Sigma (x-\mu)^2 P(X = x)} \quad \text{or} \quad \sigma = \sqrt{\Sigma x^2 P(X = x) - \mu^2}$$

- Factorial: $k! = k(k-1) \cdots 2 \cdot 1$
- Binomial coefficient: $\dbinom{n}{x} = \dfrac{n!}{x!(n-x)!}$

- Binomial probability formula

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where, n denotes the number of trials and p denotes the success probability.

- Mean of a binomial random variable: $\mu = np$
- Standard deviation of a binomial random variable: $\sigma = \sqrt{np(1-p)}$
- Poisson probability formula: $P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}$
- Mean of a Poisson random variable: $\mu = \lambda$
- Standard deviation of a Poisson random variable: $\sigma = \sqrt{\lambda}$

Discrete Random Variables and their Probability Distributions

- Mean of a discrete random variable x: $\mu = \Sigma x P(x)$
- Standard deviation of a discrete random variable x: $\sigma = \sqrt{\Sigma x^2 P(x) - \mu^2}$
- n factorial: $n! = n(n-1)(n-2) \cdots 3 \cdot 2 \cdot 1$
- Number of combinations of n items selected x at a time

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- Binomial probability formula

$$P(x) = \binom{n}{x} p^x q^{n-x}$$

- Mean and standard deviation of the binomial distribution

$$\mu = np \text{ and } \sigma = \sqrt{npq}$$

- Poisson probability formula

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

- Mean, variance, and standard deviation of the Poisson probability distribution

$$\mu = \lambda, \sigma^2 = \lambda, \text{ and } \sigma = \sqrt{\lambda}$$

Continuous Random Variables and the Normal Distribution

- z value for an x value

$$z = \frac{x - \mu}{\sigma}$$

- Value of x when μ, σ, and z are known: $x = \mu + z\sigma$

Chapter 4: Sampling Distributions

- Mean of $\bar{x}$: $\mu_{\bar{x}} = \mu$
- Standard deviation of $\bar{x}$ when $n/N \leq 0.05$: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
- z value for $\bar{x}$

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

- Population proportion: $p = X/N$
- Sample proportion: $\hat{p} = x/n$
- Mean of $\hat{p}$: $\mu_{\hat{p}} = p$
- Standard deviation of $\hat{p}$ when $n/N \leq 0.05$: $\sigma_{\hat{p}} = \sqrt{pq/n}$
- z value for $\hat{p}$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

Chapter 5: Estimation

Estimation of the Mean and Proportion

- Margin of error for the point estimation of μ: $\pm 1.96\,\sigma_{\bar{x}}$ or $\pm 1.96\,s_{\bar{x}}$

  where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and $s_{\bar{x}} = s/\sqrt{n}$
- Confidence interval for μ for a large sample

  $\bar{x} \pm z\,\sigma_{\bar{x}}$ if σ is known

  $\bar{x} \pm z\,s_{\bar{x}}$ if σ is not known

  where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ and $s_{\bar{x}} = s/\sqrt{n}$
- Confidence interval for μ for a small sample

  $\bar{x} \pm t\,s_{\bar{x}}$ where $s_{\bar{x}} = s/\sqrt{n}$
- Margin of error for the point estimation of p

  $\pm 1.96\,s_{\hat{p}}$ where $s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$
- Confidence interval for p for a large sample

  $\hat{p} \pm z\,s_{\hat{p}}$ where $s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$
- Maximum error of the estimate for μ

  $E = z\sigma_{\bar{x}}$ or $z\,s_{\bar{x}}$
- Determining sample size for estimating μ: $n = z^2\,\sigma^2/E^2$

- Maximum error of the estimate for p

  $E = z\, s_{\hat{p}}$ where $s_{\hat{p}} = \sqrt{\hat{p}\hat{q}/n}$

- Determining sample size for estimating p: $n = z^2\, pq/E^2$

Chapter 6: Hypothesis Testing

Confidence Intervals for One Population Mean

- Standardised version of the variable $\overline{x}$

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

- z-interval for μ (σ known, normal population or large sample)

$$\overline{x} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- Margin of error for the estimate of μ: $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

- Sample size for estimating μ

$$n = \left(\frac{z_{\alpha/2} . \sigma}{E}\right)^2$$

  rounded up to the nearest whole number.

- Studentised version of the variable $\overline{x}$

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$$

- t-interval for μ (σ unknown, normal population or large sample)

$$\overline{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \text{ with } df = n - 1.$$

Hypothesis Tests for One Population Mean

- z-test statistic for H_0: $\mu = \mu_0$ (σ known, normal population or large sample)

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

- t-test statistic for H_0: $\mu = \mu_0$ (σ unknown, normal population or large sample)

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \text{ with } df = n - 1.$$

Inferences for Two Population Means

- Pooled sample standard deviation

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

- Pooled t-test statistic for H_0: $\mu_1 = \mu_2$ (independent samples, normal populations or large samples, and equal population standard deviations)

$$t = \frac{\bar{x}_1-\bar{x}_2}{s_p\sqrt{(1/n_1)+(1/n_2)}} \text{ with } df = n_1+n_2-2.$$

- Pooled t-interval for $\mu_1 - \mu_2$ (independent samples, normal populations or large samples, and equal population standard deviations)

$$(\bar{x}_1-\bar{x}_2) \pm t_{\alpha/2}\cdot s_p\sqrt{(1/n_1)+(1/n_2)} \text{ with } df = n_1+n_2-2.$$

- Degrees of freedom for non-pooled t-procedures

$$\Delta = \frac{[(s_1^2/n_1)+(s_2^2/n_2)]^2}{\frac{(s_1^2/n_1)^2}{n_1-1}+\frac{(s_2^2/n_2)^2}{n_2-1}} \text{ rounded down to the nearest integer.}$$

- Non-pooled t-test statistic for H_0: $\mu_1 = \mu_2$ (independent samples, and normal populations or large samples)

$$t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{(s_1^2/n_1)+(s_2^2/n_2)}} \text{ with } df = \Delta.$$

- Non-pooled t-interval for $\mu_1 - \mu_2$ (independent samples, and normal populations or large samples):

$$(\bar{x}_1-\bar{x}_2) \pm t_{\alpha/2}\cdot\sqrt{(s_1^2/n_1)+(s_2^2/n_2)} \text{ with } df = \Delta.$$

- Paired t-test statistic for H_0: $\mu_1 = \mu_2$ (paired sample, and normal differences or large sample)

$$t = \frac{\bar{d}}{s_d/\sqrt{n}} \text{ with } df = n-1.$$

- Paired t-interval for $\mu_1 - \mu_2$ (paired sample, and normal differences or large sample)

$$\bar{d} \pm t_{\alpha/2}\cdot\frac{s_d}{\sqrt{n}} \text{ with } df = n-1.$$

Inferences for Population Standard Deviations

- χ^2-test statistic for H_0: $\sigma = \sigma_0$ (normal population)

$$\chi^2 = \frac{n-1}{\sigma_0^2}s^2 \text{ with } df = n-1.$$

- χ^2-interval for a (normal population)

$$\sqrt{\frac{n-1}{\chi^2_{\alpha/2}}}.s \text{ or } \sqrt{\frac{n-1}{\chi^2_{1-\alpha/2}}}.s \text{ with } df = n-1.$$

- F-test statistic for H_0: $\sigma_1 = \sigma_2$ (independent samples and normal populations)

$$F = s_1^2 / s_2^2 \text{ with } df = (n_1 - 1, n_2 - 1).$$

Inferences for Population Proportions

- Sample proportion

$$\hat{p} = \frac{x}{n}$$

where x denotes the number of members in the sample that have the specified attribute.

- z-interval for p

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}$$

(*Assumption:* both x and $n - x$ are 5 or greater)

- Margin of error for the estimate of p

$$E = z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}$$

- Sample size for estimating p

$$n = 0.25\left(\frac{z_{\alpha/2}}{E}\right)^2 \quad \text{or} \quad n = \hat{p}_g(1-\hat{p}_g)\left(\frac{z_{\alpha/2}}{E}\right)^2$$

rounded up to the nearest whole number (g = "educated guess")

- z-test statistic for H_0: $p = p_0$

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

(*Assumption:* both np_0 and $n(1 - p_0)$ are 5 or greater)

- Pooled sample proportion: $\hat{p}_p = \dfrac{x_1 + x_2}{n_1 + n_2}$

- z-test statistic for H_0: $p_1 = p_2$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p(1-\hat{p}_p)}\sqrt{(1/n_1) + (1/n_2)}}$$

(*Assumptions:* independent samples; x_1, $n_1 - x_1$, x_2, $n_2 - x_2$ are all 5 or greater)

- z-interval for $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \cdot \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

(*Assumptions:* independent samples; x_1, $n_1 - x_1$, x_2, $n_2 - x_2$ are all 5 or greater)

- Margin of error for the estimate of $p_1 - p_2$

$$E = z_{\alpha/2} \cdot \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

- Sample size for estimating $p_1 - p_2$

$$n_1 = n_2 = 0.5\left(\frac{z_{\alpha/2}}{E}\right)^2$$

or

$$n_1 = n_2 = (\hat{p}_{1g}(1-\hat{p}_{1g}) + \hat{p}_{2g}(1-\hat{p}_{2g}))\left(\frac{z_{\alpha/2}}{E}\right)^2$$

rounded up to the nearest whole number (g = "educated guess")

Chapter 7: Curve Fitting, Regression and Correlation

- Simple linear regression module for population
  $y = A + Bx + \varepsilon$
- Estimated simple linear regression model

$$\hat{y} = a + bx$$

- S_{xx}, S_{xy}, and S_{yy}

$$SS_{xx} = \Sigma(x_i - \bar{x})^2 = \Sigma x_i^2 - (\Sigma x_i)^2 / n$$

$$SS_{xy} = \Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i) / n$$

$$SS_{yy} = \Sigma(y_i - \bar{y})^2 = \Sigma y_i^2 - (\Sigma y_i)^2 / n$$

- Regression equation: $\hat{y} = a + bx$

where $$b = \frac{SS_{xy}}{SS_{xx}}, \quad a = \bar{y} - b\bar{x} = \frac{1}{n}(\Sigma y_i - b\Sigma x_i)$$

- Total sum of squares: $\text{SST} = \Sigma(y_i - \bar{y})^2 = SS_{yy}$
- Regression sum of squares: $\text{SSR} = \Sigma(\hat{y}_i - \bar{y})^2 = SS_{xy}^2 / SS_{xx}$
- Error sum of squares: $\text{SSE} = \Sigma(y_i - \hat{y}_i)^2 = SS_{yy} - SS_{xy}^2 / SS_{xx}$
- Regression identity: SST = SSR + SSE
  or SSR = SST – SSE
- Coefficient of determination: $r^2 = \dfrac{\text{SSR}}{\text{SST}} = b\dfrac{SS_{xy}}{SS_{yy}}$
- Linear correlation coefficient

$$r = \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \quad \text{or} \quad r = \frac{SS_{xy}}{\sqrt{SS_{xx} SS_{yy}}}$$

Inferential Methods in Regression and Correlation

- Population regression equation: $y = a + bx$
- Standard error of the estimate

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}}$$

- Test statistic for a test of hypothesis about B

$$t = \frac{b - B}{s_b}$$

where $$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

$$s_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}}$$

- Test statistic for H_0: $\beta_1 = 0$

$$t = \frac{b_1}{s_e / \sqrt{SS_{xx}}} \text{ with } df = n - 2$$

- Confidence interval for β_1

$$b_1 \pm t_{\alpha/2} \cdot \frac{s_e}{\sqrt{SS_{xx}}} \text{ with } df = n - 2$$

- Confidence interval for the conditional mean of the response variable corresponding to x_p

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_e \sqrt{\frac{1}{n} + \frac{(x_p - \Sigma x_i / n)^2}{SS_{xx}}} \text{ with } df = n - 2$$

- Confidence interval for $\mu_{y|x}$

$$\hat{y} \pm t\, s_{\hat{y}_m}$$

where $$s_{\hat{y}_m} = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

- Prediction interval for y_p

$$\hat{y} \pm t\, s_{\hat{y}_p}$$

where $$s_{\hat{y}_p} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_{xx}}}$$

- Prediction interval for an observed value of the response variable corresponding to x_p

$$\hat{y}_p \pm t_{\alpha/2} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \Sigma x_i / n)^2}{SS_{xx}}} \text{ with } df = n - 2$$

- Test statistic for H_0: $\rho = 0$

$$t = \frac{r}{\sqrt{\dfrac{1 - r^2}{n - 2}}} \text{ with } df = n - 2$$

- Test statistic for a correlation test for normality

$$R_p = \frac{\Sigma x_i w_i}{\sqrt{SS_{xx}\Sigma w_i^2}}$$

where x and w denote observations of the variable and the corresponding normal scores, respectively.

Chapter 8: chi-square Tests

- Expected frequency for a category for a goodness-of-fit test

$$E = np$$

- Degrees of freedom for a goodness-of-fit test

$$df = k - 1$$

where k is the number of categories.

- Expected frequency for a cell for an independence or homogeneity test

$$E = \frac{\text{(row total)(column total)}}{n}$$

- Degrees of freedom for a test of independence or homogeneity

$$df = (R-1)(C-1)$$

where R and C are the total number of rows and columns, respectively, in the contingency table.

- Test statistic for a goodness-of-fit test and a test of independence or homogeneity

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- Confidence interval for the population variance σ^2

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2}} \text{ to } \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}$$

- Test statistic for a test of hypothesis about σ^2

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Chapter 9 : Analysis of Variance

Let k = the number of different samples (or treatments)
n_i = the size of sample i
T_i = the sum of the values in sample i
n = the number of values in all samples $= n_1 + n_2 + n_3 + \cdots$
Σx = the sum of the values in all samples $= T_1 + T_2 + T_3 + \cdots$
Σx^2 = the sum of the squares of values in all samples

- For the F-distribution

  Degrees of freedom for the numerator $= k - 1$
  Degrees of freedom for the denominator $= n - k$

- Between-samples sum of squares

$$\text{SSB} = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right) - \frac{(\Sigma x)^2}{n}$$

- Within-samples sum of squares

$$\text{SSW} = \Sigma x^2 - \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \cdots\right)$$

- Total sum of squares

$$\text{SST} = \text{SSB} + \text{SSW} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

- Variance between samples

$$\text{MSB} = \frac{\text{SSB}}{k-1}$$

- Variance within samples

$$\text{MSW} = \frac{\text{SSW}}{n-k}$$

- Test statistic for a one-way ANOVA test

$$F = \frac{\text{MSB}}{\text{MSW}}$$

Answers to Selected Problems

CHAPTER–1

1.1 *(a)* and *(b)* See Table 1.1.

Table 1.1

Category	Frequency	Relative frequency	Percentage
X	10	10/30 = 0.3333	33.33
Y	9	9/30 = 0.3	30
Z	11	11/30 = 0.3667	36.67
Total	30	1.0	100

(c) 33%

(d) 70%

(e) See Fig. 1.1.

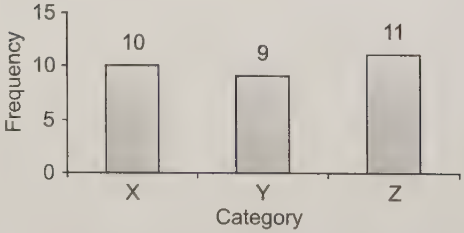


Fig. 1.1: Bar graph for the frequency distribution

1.2 *(a)* and *(b)* See Table 1.2.

Table 1.2

Category	Frequency	Relative frequency	Percentage
F	6	6/15 = 0.4	40
A	5	5/15 = 0.3333	33.33
N	4	4/15 = 0.2667	26.67
Total	15	1.0	100

(c) 40% of the faculty members are in favour of this issue.

1.3 *(d)*

Weight (kg)	Tally	f
72–74	||||	4
75–77	𝍸 |	6
78–80	𝍸 ||||	9
81–83	𝍸 ||	7
84–86	||||	4
Total	30	30

1.4

Vehicles owned	Number of households (f)
0	2
1	18
2	12
3	3
4	4
5	1
Total	$\Sigma f = 40$

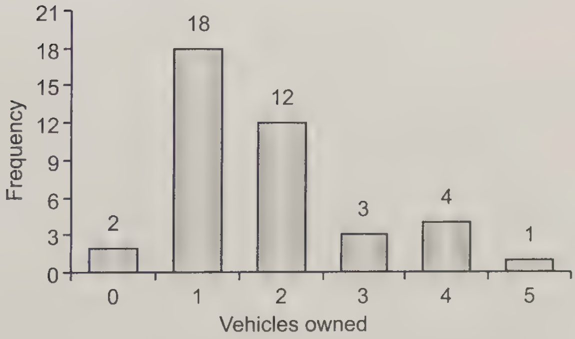


Fig. 1.4: Bar graph vehicles owned

1.5 *(a)* and *(c)* – see Table 1.5

Table 1.5

Class boundaries	Class mid point	Frequency	Relative frequency	Percentage
17.5 to less than 30.5	24	12	0.24	24
30.5 to less than 43.5	37	18	0.36	36
43.5 to less than 56.5	50	14	0.28	28
56.5 to less than 69.5	63	6	0.12	12
Total		50	1.0	100

(*b*) Yes, each class has a width of 13.

(*d*) 24 + 36 = 60% of the employees is 43 years old or younger.

1.6 (*a*) and (*b*) See Table 1.6

Table 1.6

Number of machines manufactured	Frequency	Relative frequency	Percentage
21-23	7	0.233	23.3
24-26	4	0.133	13.3
27-29	8	0.2667	26.67
30-32	5	0.1667	16.67
33-35	6	0.2	20
Total	30	1.0	100

(*c*) See Fig. 1.6

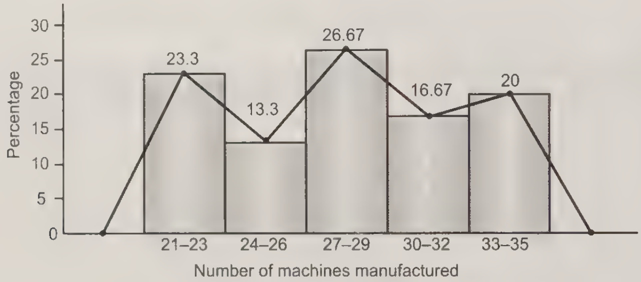


Fig. 1.6

(*d*) For 26.67% of the days the number of machines manufactured is in the interval 27 to 29.

1.7 (*a*) and (*b*)

Table 1.7

Category	Frequency	Relative frequency	Percentage
F	7	7/15 = 0.467	46.7
A	5	5/15 = 0.333	33.3
AS	3	3/15 = 0.2	20.0
Total	15	1.0	100

(*c*) 46.7% of the faculty in this sample is full professors.

(*d*) See Fig. 1.7.

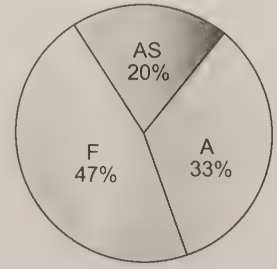


Fig. 1.7

1.8 *(a)* and *(b)*: The frequency distribution are given in Table 1.8.

Table 1.8

Category	Frequency	Relative frequency	Percentage
1	8	0.308	30.8
2	9	0.346	34.6
3	9	0.346	34.6
Total	26	1.0	100

(c) 34.6% of the elements in this sample belong to category 2.

(d) 30.8% + 34.6% = 65.4% of the elements belong to categories 1 and 3 in this sample.

1.9 *(a)* and *(b)*: The frequency distribution, relative frequencies and percentage for all classes are shown in Table 1.9.

Table 1.9

Number of oil changes	Frequency	Relative frequency	Percentage
21-23	5	0.167	16.7
24-26	5	0.167	16.7
27-29	9	0.3	30
30-32	6	0.2	20
33-35	5	0.167	16.7
	30	1.0	100

(c) The histogram and the polygon for the percentage distribution are shown in Fig.1.9.

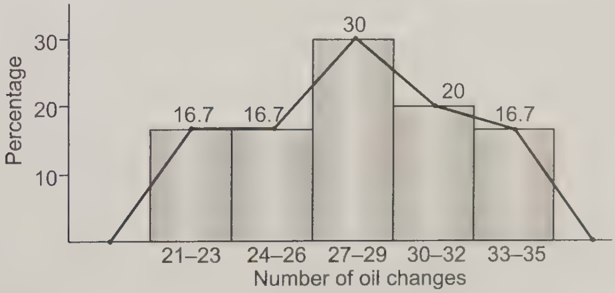


Fig. 1.9: Histogram and polygon for the percentage distribution

(d) 20% of the days the oil changes were made in the interval 30–32.

1.10 *(a)* The cumulative frequency, cumulative relative frequency and cumulative percentage distribution table are presented in Table 1.10.

Table 1.10

Collision claims $	Cumulative frequency	Cumulative relative frequency	Cumulative percentage
1–1000	7	0.292	29.2
1–2000	15	0.625	62.5
1–3000	21	0.875	87.5
1–4000	22	0.917	91.7
1–5000	24	1.0	100

(*b*) The ogive for the cumulative percentage distribution is shown in Fig.1.10.

(*c*) From Fig.1.10, the percentage of claimants in the sample who received $2500 or less is approximately 78%.

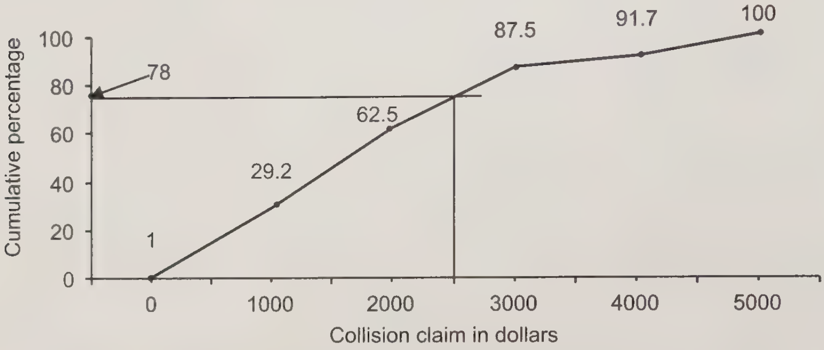


Fig. 1.10: Ogive for the cumulative percentage distribution

1.11 (*a*) and (*b*): Refer to the Table 1.11 given below:

Table 1.11

Automobile repair costs (dollars)	Frequency	Relative frequency	Percentage
1–1500	10	0.4	40
1501–3000	9	0.36	36
3001–4500	1	0.04	4
4501–6000	5	0.2	20
	25	1.0	100

(*c*) See Fig. 1.11.

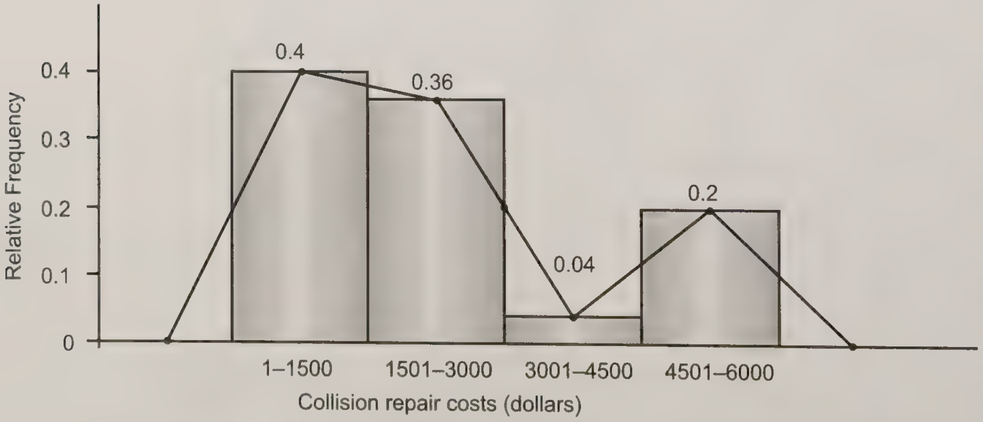


Fig. 1.11: Relative frequency versus collision repair costs (dollars)

(*d*) The class boundaries of the fourth class are $4500.50 and $6000.50. The width of the class is $1500.

1.12 (*a*) and (*b*): See Table 1.12.

Table 1.12

Number of unemployed adults	Frequency	Relative frequency	Percentage
1–70	10	0.50	50
71–140	4	0.20	20
141–210	1	0.05	5
211–280	2	0.10	10
281–350	3	0.15	15
Total	20	1.0	100

1.13 *(a)* and *(b)*: See Table 1.13.

Table 1.13

Annual compensation (thousands of dollars)	Frequency	Relative frequency	Percentage
1 to less than 1,000	7	0.241	24.1
1,000 to less than 2,000	10	0.345	34.5
2,000 to less than 3,000	8	0.276	27.6
3,000 to less than 4,000	1	0.034	3.4
4,000 to less than 5,000	3	0.104	10.4
Total	29	1.0	100

1.14 *(a)* and *(b)*: See Table 1.14.

Table 1.14

Number of children less than 18 years age	Frequency	Relative frequency	Percentage
0	9	0.30	30
1	8	0.266	26.6
2	11	0.367	36.7
3	2	0.067	6.7
Total	30	1.0	100

(c) 11 + 2 = 13 families have 2 or 3 children less than 18 years of age.

(d) See Fig. 1.14.

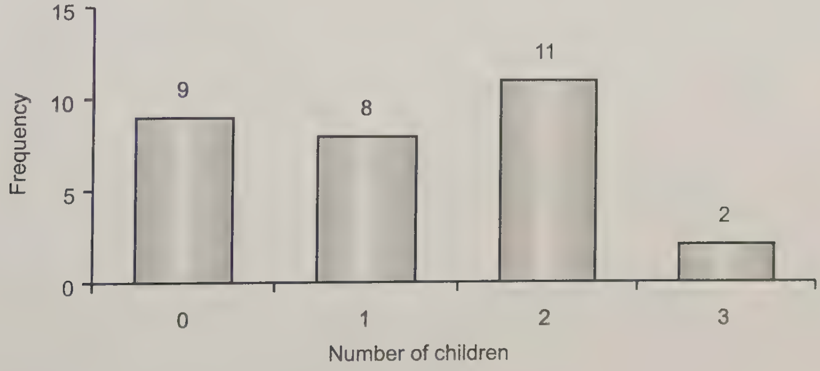


Fig. 1.14

1.15 Refer to the solution of problem 1.5.

(a) and *(b)*: See Table 1.15.

Table 1.15

Age (years)	Cumulative frequency	Cumulative relative frequency	Cumulative percentage
18 to 30	12	0.24	24
18 to 43	30	0.60	60
18 to 56	44	0.88	88
18 to 69	50	1.00	100

(c) 100 – 60 = 40% of the employees are 44 years of age or older.

(d) See Fig. 1.15.

(e) About 52% of the employees are 40 years old or younger.

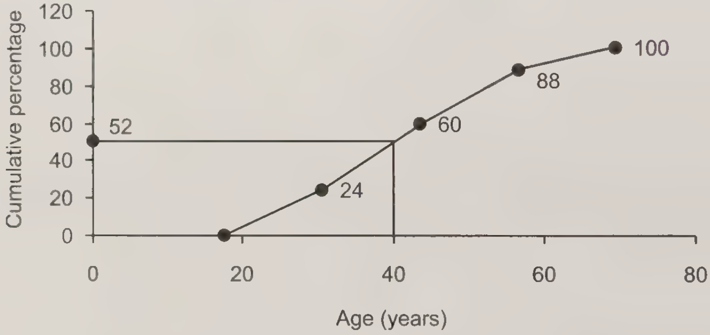


Fig. 1.15

1.16 Refer to the solution of problem 1.6.

Table 1.16

Number of machines manufactured	Cumulative frequency	Cumulative Relative frequency	Cumulative percentage
21–23	7	0.233	23.3
24–26	11	0.367	36.7
27–29	19	0.634	63.4
30–32	24	0.800	80
33–35	30	1.0	100

1.17

Number of machines manufactured	Cumulative frequency	Cumulative relative frequency	Cumulative percentage
1 to less than 1,000	7	0.241	24.1
1,000 to less than 2,000	17	0.586	58.6
2,000 to less than 3,000	25	0.862	86.2
3,000 to less than 4,000	26	0.896	89.6
4,000 to less than 5,000	29	1.0	100

1.18

7	86	73	
8	92	81	
9	97	95	80
10	42	84	66
11	70	28	

Stem and leaf display

7	73	86	
8	81	92	
9	80	95	97
10	42	66	84
11	28	70	

Stem and leaf display in increasing order

1.19

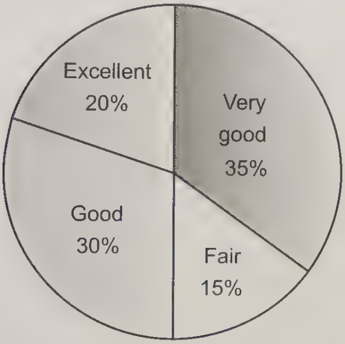
0	5	7				
1	0	2	7	3	9	
2	3	7	6	9	1	2
3	4	7	2			
4	7	3				
5	0					
6	4					

Stem and leaf display

0	5	7				
1	0	2	3	7	9	
2	1	2	3	6	7	9
3	2	4	7			
4	3	7				
5	0					
6	4					

Stem and leaf display in increasing order

1.20



Excellent	80	$\frac{80}{400} = 20\%$
Very good	140	$\frac{140}{400} = 35\%$
Good	120	$\frac{120}{400} = 30\%$
Fair	60	$\frac{60}{400} = 15\%$
Σ	400	100%

1.21

Stem	Leaves																
4	3	6	8														
5	0	1	1	2	2	3	3	5	5	6	6	7	8				
6	0	6	7	7	8	8											
7	0	1	2	3	3	4	5	5	5	5	7	7	8	8			
8	0	0	0	0	1	2	2	2	3	3	4	4	5	5	6	6	6
9	1	2	3														

1.22

Stem	Leaves			
18	4	6		
19	5	0	7	8
20	4	5		
21	1	4	5	6
22	5	4		
23	0	9	7	3
24	1			
25	4			

1.23

Stem	Leaves					
5	0	3	8			
6	2	4	6	8	9	
7	1	2	2	2	5	6
8		7	8	9		
9	0	1	3	4	6	7
		7				
	2	3	3	5	6	9

1.24

Stem	Leaves					
4	28					
5	86	50	40	78		
6	75	20	60	51		
7	33	50	65	81	50	
8	71	80	40	70	01	20
9	89	56	30	75	50	∞
10	21	30	70	50		

1.25 *(a)*

Stem-and-leaf display

0	89										
1	73	09	78	17	89	75					
2	81	87	35	87							
3	17	79	55	54	57	88					
4	73	12	53	17	74	54	89	84	14	69	
5	89										
6	08	16	11	01							

(b)

Stem-and-leaf display by group

0–1	89	*	73		09		78		17		89		75				
2–3	81	87	35		87	*	17	79	55	54	57	88					
4–6	73	12	53	17	74	54	89	84	14	69	*	89	*	08	16	11	01

1.26 Mean = 5.6 hours, median = 6 and mode 0 and 6
1.27 Mean = 25.28, median = 25.8 and mode = 24
1.28 Mean = 78.42, median = 77.0 and no mode
1.29 Mean = 5.70, median = 5.7 and mode = 5.7
1.30 Mean = 27.5, median = 27.5 and mode = 23
1.31 Range = 11, variance = 10.788 and standard deviation = 3.28
1.32 Range = 13, variance = 19.822 and standard deviation = 4.452
1.33 Range = 26, variance = 72.278 and standard deviation = 8.50
1.34 Range = 12, mean = 9.4, variance = 11.4 and standard deviation = 3.3764
1.35 Variance = 113.3444 and standard deviation = 10.6463
1.36 Range = 18, variance = 39.5667 and standard deviation = 6.2902
1.37 Standard deviation: Data set 1 = 14.64 and Data set 2 = 14.64
1.38 Mean = 67.5, variance = $2,432.20 and standard deviation = $49.32
1.39 Mean = 9.5897, variance = 39.7256 and standard deviation = 6.3028
1.40 Mean = 28.4, variance = 210.6531 and standard deviation = 14.5139
1.41 Mean = 3.4, variance = 0.8990 and standard deviation = 0.9482
1.42 Mean = 24.65, variance = 122.7462 and standard deviation = 11.08
1.43 Mean = 1.7875, variance = 1.0049 and standard deviation = 1
1.44 *(a)* $Q_1 = 69.5, Q_2 = 76.5, Q_3 = 85.5, IQR = 16$. The value of 80 lies between Q_2 and Q_3.
  (b) $P_{93} = 90$
  (c) percentile rank of 81 = 66.67%
1.45 *(a)* $Q_1 = 35, Q_2 = 40, Q_3 = 43, IQR = 8$
  (b) $P_{79} = 45$
  (c) Percentile rank of 39 = 36.67%
1.46 *(a)* $Q_1 = 24, Q_2 = 27.5, Q_3 = 31, IQR = 7$
  (b) Less than 29 for approximately 65% of the days.
  (c) 76.67%
1.47 *(a)* $Q_1 = 69, Q_2 = 78, Q_3 = 84, IQR = 15$; 81 lies between Q_2 and Q_3
  (b) $P_{93} = 97$
  (c) 57.89% of the scores were below 82
1.48 *(a)* $Q_1 = 73.5, Q_2 = 79, Q_3 = 80$
  (b) $IQR = 6.5$
  (c) Min = 46, $Q_1 = 73.5, Q_2 = 79, Q_3 = 80$ and Max = 83
1.49 *(a)* $Q_1 = 4, Q_2 = 7, Q_3 = 12$
  (b) $IQR = 8$
  (c) Min = 1, $Q_1 = 4, Q_2 = 7, Q_3 = 12$ and Max = 25
1.50 *(a)* $Q_1 = 70.5, Q_2 = 77.5$, and $Q_3 = 86.5$
  (b) $IQR = 16$
1.51 *(a)* $Q_1 = 30.5, Q_2 = 37$, and $Q_3 = 49$
  (b) $IQR = 18.5$

1.52 $P_{62} = 79.5$

1.53 $P_{85} = 66.67\%$

1.54 $Q_1 = 23$, $Q_2 = 30.5$, and $Q_3 = 36.5$ hours.

1.55 *(a)* Min = 6, $Q_1 = 23$, $Q_3 = 36.5$, Max = 64 hours.

(b) IQR = 13.5 hours.

(c) Lower limit = 2.75 hours; Upper limit = 56.75 hours.

(d) 64 is a potential outlier.

1.57 *(a)* $Q_1 = 88$, $Q_2 = 131.5$ and $Q_3 = 154$

(b) IQR = 66

(c) No potential outliers

(d)

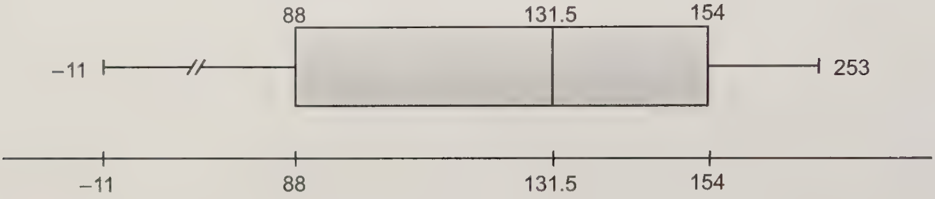
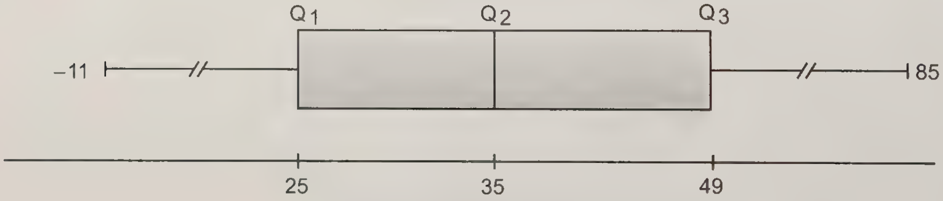


Fig. 1.57: Box plot

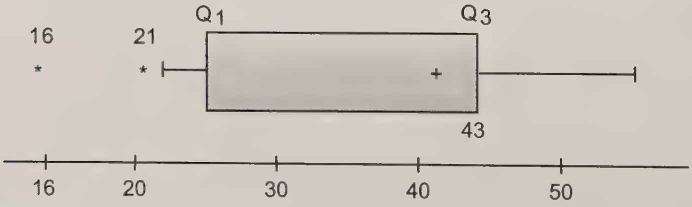
1.58 Skewed to the left.

1.59



Box-and-whisker plot

1.60



Box-and-whisker plot

1.61 The data are nearly symmetric.

CHAPTER – 2

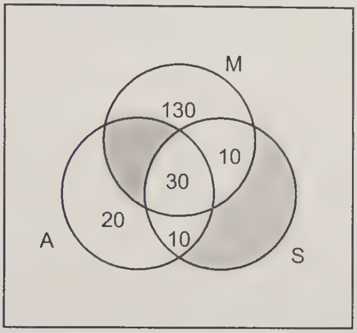
2.1 *(a)* needs engine-oil change, tune-up and new tires

(b) needs tune-up, new tires but no engine-oil change

(c) needs engine-oil change but neither tune-up nor new tires
(d) need engine-oil change and new tires
(e) need tune-up, but no new tires
(f) does not need oil change.

2.2 *(a)* 5
(b) 1 and 2 together
(c) 3, 5 and 6 together
(d) 1, 3, 4 and 6 together

2.3 *(a)*



(b) M has 170 systems, S has 50 systems

2.4 *(a)* $P(S) = 0.56$
(b) $S = (A \text{ or } B \text{ or } C)$
(c) $P(A) = 0.01, P(B) = 0.14, P(C) = 0.41$
(d) $P(S) = 0.56$

2.5 *(a)* 0.99
(b) 0.56

2.6 *(a)* $P(F) = 0.50$; $P(D) = 0.10$; $P(F \text{ and } D) = 0.05$
(b) $P(F \text{ or } D) = 0.55$
(c) $P(M) = 0.50$

2.7 0.96875

2.8 $S = \{MM, MF, FM, FF\}$

2.9 *(a)* FF
(b) $\{FF, FA, AF\}$ *(c)* $\{FA, AF\}$

2.10 0.3334

2.11 0.1667

2.12 *(a)* 1/8
(b) Sample space and sample points
1, 1 2, 1 3, 1 4, 1 5, 1 6, 1
1, 2 2, 2 3, 2 4, 2 5, 2 6, 2
1, 3 2, 3 3, 3 4, 3 5, 3 6, 3

1,4 2,4 3,4 4,4 5,4 6,4
1,5 2,5 3,5 4,5 5,5 6,5
1,6 2,6 3,6 4,6 5,6 6,6
Probability = 1/36

(c) (i) 3/8 (ii) 1/2
2.13 (a) (i) 1/6 (ii) 1/12 (b) 2/3
2.14 (a) 0.2125
(b) 0.0962
(c) 21.25% of the employees are in their 50's; 9.62% of the rank R_3 employees are in their 50's
2.15 $P(R_3 \text{ and } A_4) = 0.2655$; $P(R_3) = 0.2706$; $P(A_4/R_3) = 0.0962$
2.16 (a) $P(M_4/S_1) = 0.0888$
(b) $P(S_1/M_4) = 0.6666$
2.17 (a) $P(M_2/F_1) = 0.6122$; $P(F_1 \text{ and } M_2) = 0.2449$
2.18 0.0316
2.19 0.075
2.20 0.9688
2.21 (a) 0.001
(b) 0.999
(c) The intersection of the events *A*, *B* and *C* is denoted as *ABC* in Fig. 2.21.

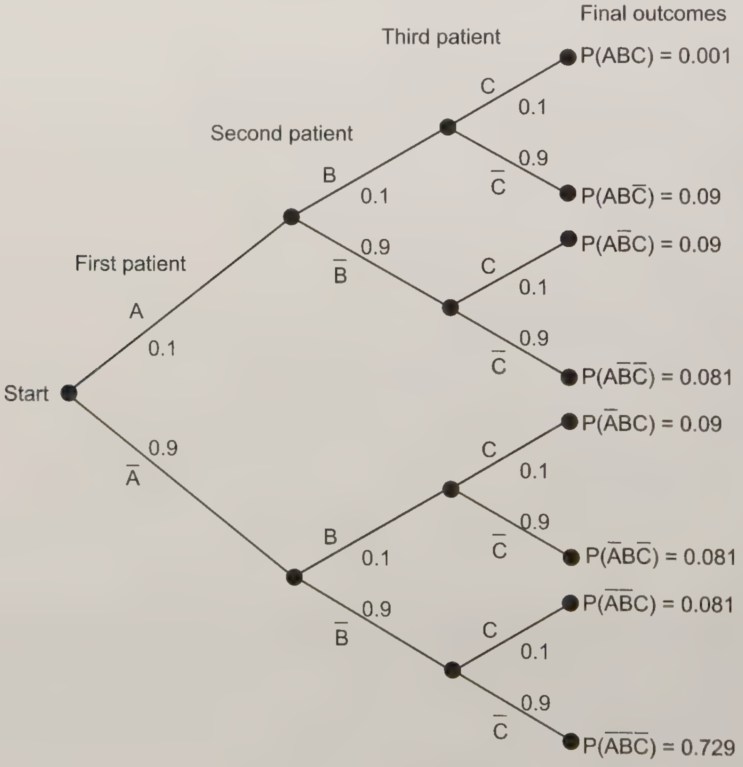


Fig. 2.21: Tree diagram

2.22 $S = \{MMM, MMW, MWM, MWW, WMM, WMW, WWM, WWW\}$

2.23 $P(\text{woman}) = 0.4$; $P(\text{man}) = 0.6$; Sum of the probabilities = 1

2.24 *(a)* See Table 2.24 (Frequency distribution).

Table 2.24

Number of design engineers	Frequency	Relative frequency
0	35	0.035
1	210	0.210
2	379	0.379
3	270	0.270
4 or more	106	0.106
	1000	1.0

(b) 0.379
(c) 0.106

2.25 *(a)* not independent
(b) not mutually exclusive

2.26 *(a)* 3/13
(b) 5/8

2.27 *(a)* 0.21
(b) 0.24
(c) not independent

2.28 *(a)* 23/40

2.29 *(a)* 0.156
(b) 0.1265

2.30 *(a)* *(i)* 0.025 *(ii)* 0.183 *(b)* 0

2.31 0.2651

2.32 0.6667

2.33 *(a)* 0.5428
(b) 0.8437

2.34 0.9516

2.35 *(a)* mutually exclusive
(b) not mutually exclusive

2.36 $P(\bar{B}) = 0.003$; $P(B) = 0.097$; $P(C) = 0.0941$; $P(D) = 0.9941$; $P(\bar{D}) = 0.0059$; $P(\bar{C}) = 0.0029$

2.38 *(a)* 5/8
(b) 0.4

2.39 0.486

2.40 0.9882

2.41 50%

2.42 *(a)* 0.029
(b) 0.3103

2.43 7/10

2.44 6/100

2.45 *(a)* $S = \{FF, FA, AF, AA\}$

(b) 0.84

2.46 0.135

2.47 *(a)* 0.04

(b) 0.96

(c) 0.5

(d) 0.33

2.48 *(a)* 1/3

(b) $P(A) = 5/9$, $P(B/A) = 1/2$, $P(AB) = 5/18$

2.49 0.4437

2.50 0.99992

2.51 *(a)* 0.4545

(b) 0.2424

(c) 0.3030

2.52 0.3182

2.53 *(a)* 0.25

(b) 0.36

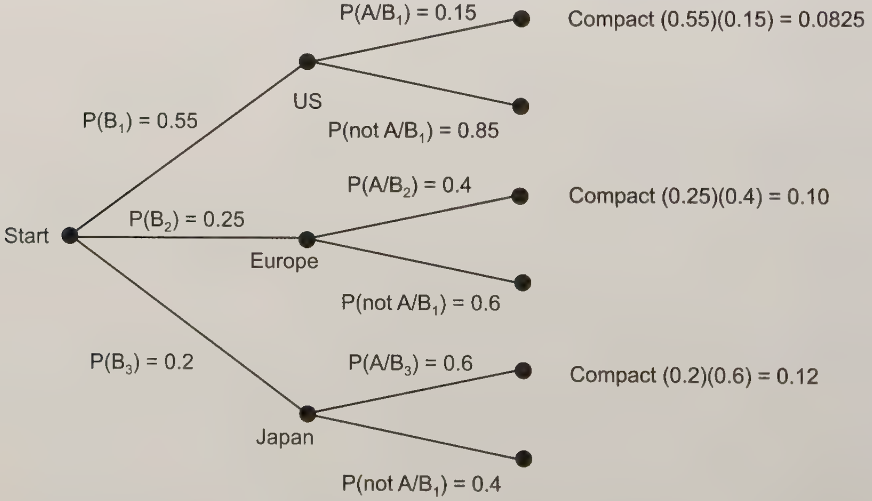
2.54 0.7308

2.55 *(a)* 1/3

(b) 2/3

2.56 *(a)* 0.3025

(b) A tree diagram showing the various probabilities along with the probabilities as branches is shown in Fig. 2.56.



$P(A/B_1) = 0.15$
Compact (0.55)(0.15) = 0.0825
US
$P(B_1) = 0.55$
$P(\text{not } A/B_1) = 0.85$
$P(A/B_2) = 0.4$
Compact (0.25)(0.4) = 0.10
$P(B_2) = 0.25$
Start
Europe
$P(\text{not } A/B_1) = 0.6$
$P(B_3) = 0.2$
$P(A/B_3) = 0.6$
Compact (0.2)(0.6) = 0.12
Japan
$P(\text{not } A/B_1) = 0.4$

Fig. 2.56: Tree diagram

(c) 0.33058

2.57 (a) 0.595

(b) 0.605

(c) 0.0987

2.58 (a) 0.65

(b) 0.23077

(c) Here, we like to find $P(d_3/A)$, by applying the Baye's formula.

$$P(d_3/A) = \frac{P(A/d_3)\cdot P(d_3)}{P(A)} = \frac{(0.6)(0.25)}{0.65} = 0.23077$$

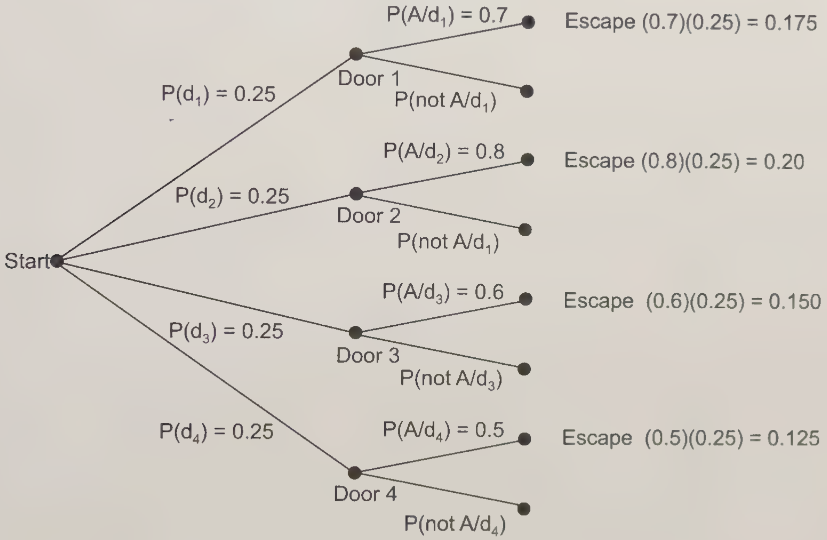


Fig. 2.58: Tree diagram

2.59 0.2384

2.60 0.2857

CHAPTER – 3

3.1 (a) Table 3.1 gives the probability distribution.

Table 3.1: Probability distribution

x	P(x)
2	11/90 = 0.12
3	20/90 = 0.22
4	23/90 = 0.25
5	22/90 = 0.24
6	14/90 = 0.15

(b) approximate (c) (i) 0.22, (ii) 0.86, (iii) 0.59, (iv) 0.34

3.2 *(a)* 0.12
(b) 0.41
(c) 0.43
(d) 0.36

3.3 $\mu = 3.25; \sigma = 1.26$

3.4 *(a)* $\mu = 1.30$
(b) $\sigma^2 = 1.51; \sigma = 1.23$

3.5 *(a)*

Table 3.5(a)

x	P(x)
0	0.15
1	0.20
2	0.35
3	0.30
Σ	1.00

Figure 3.5: shows the line graph of the probability distribution of Table 3.5(a).

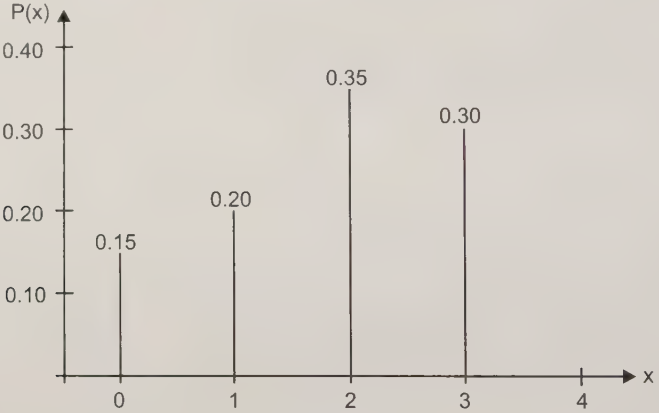


Fig. 3.5

(b) *(i)* 0.35, *(ii)* 0.70, *(iii)* 0.65, *(iv)* 0.35

3.6 *(a)* Discrete random variable
(b) Continuous random variable
(c) Discrete random variable
(d) Continuous random variable
(e) Continuous random variable
(f) Discrete random variable
(g) Discrete random variable

3.7 *(a)* Yes
(b) 0.5
(c) 0.45

3.8 $\sigma = 1.20$

3.9 $\mu = 12.5; \sigma = 1.36$

3.10 $\mu = 3.5; \sigma^2 = 2.92$

3.11 0.39347

3.12 900 hours

3.13

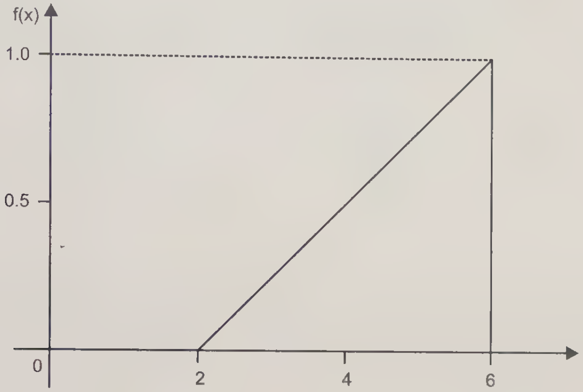


Fig. 3.13(a): Density function

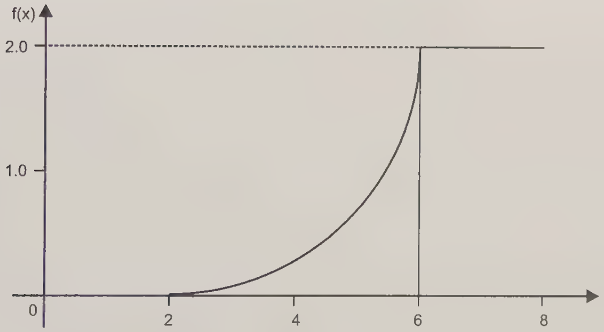


Fig. 3.13(b): Distribution function

3.14 *(a)* 0.5941

(b) 0.1992

3.15 *(a)* ¼

(b) 39/64

(c) 1/16

3.16 0.0819

3.17 $\mu = 0.4167; \sigma^2 = 0.0645$

3.18 *(a)* $\mu = (a + b)/2; \sigma^2 = (b - a)^2/12$

3.19 $x_m = 6.931$ sec.

3.20 $\mu = \lambda$; variance $= \lambda^2$

3.21 *(a)* 26!
(b) 720

3.22 315

3.23

Table 3.23: Number of possible acronyms with *n* letters

n	Number of possibilities with repetition	Number of possibilities without repetition
2	676	650
3	17,576	15,600
4	456,976	358,800
5	11,881,376	7,893,600

3.24 *(a)* 90,000
(b) 27,216
(c) The first digit can be chosen in 9 ways, the second in 8 ways, the third in 7 ways and the fourth in 6 ways. Then 9.8.7.6 = 3,024 numbers can be formed.

3.25 3003

3.26 *(a)* 0.4118
(b) 0.068
(c) 0.0229
(d) 0.7941

3.27 120

3.28 12

3.29 24

3.30 7/10

3.31 *(a)* 0.57664
(b) 0.35161
(c) 4.72×10^{-7}

3.32 *(a)* 462
(b) 462

3.33 *(a)* 2520
(b) 1400
(c) 504

3.34 210

3.35 *(a)* 350
(b) 150
(c) 105

3.36 *(a)* 140,608
(b) 132,600

3.37		56
3.38		4
3.39		6
3.40		10
3.41	*(a)*	$P(X \geq 1) = 0.9576$
	(b)	$P(X \geq 1) = 0.5429$
3.42		0.44
3.43		$P(1) = 0.347$; $P(2) = 0.069$; $P(3) = 0.005$; $P(0) = 0.579$
3.44		0.2759
3.45	*(a)*	$n = 50$ and $p = 0.1$
	(b)	0.112
	(c)	4.51×10^{-48}
3.46		$P(\text{return}) = 0.19$
3.47		0.0086
3.48	*(a)*	0.228
	(b)	0.9655
3.49	*(a)*	0.00356
	(b)	0.0444
3.50		0.122
3.51	*(a)*	0.4575
	(b)	0.082
3.52		0.423
3.53	*(a)*	0.1755
	(b)	0.0858
	(c)	0.2873
3.54	*(a)*	0.2240
	(b)	0.1992
3.55		$P(X \leq 3) = 0.433$ (using Poisson distribution with $\lambda = 4$)
3.56		0.453
3.57	*(a)*	0.0041
	(b)	0.0467
	(c)	0.2765
	(d)	0.8203
3.58	*(a)*	0.4096
	(b)	0.4096
	(c)	0.1536
3.59	*(a)*	0.18
	(b)	0.323

3.60	*(a)*	$2000\,C_3(0.001)^3(0.999)^{1997}$
	(b)	$1-[2000\,C_0(0.001)^0(0.999)^{2000}+2000\,C_1(0.001)^1(0.999)^{1999}+2000\,C_2(0.001)^2(0.949)^{1998}]$
3.61		0.228
3.62		21.12%
3.63		11.48%
3.64		0.9772
3.65		0.8185
3.66	*(a)*	0.06681
	(b)	0.86638
	(c)	$\sigma = 0.000214$
3.67	*(a)*	0.8849
	(b)	0.1151
	(c)	0.7262
3.68	*(a)*	0.6772
	(b)	0.6772
	(c)	0.0668
	(d)	0.2560
3.69	*(a)*	0.99865
	(b)	0.98422
	(c)	0.9973
	(d)	0.34134
	(e)	0.00135
3.70	*(a)*	0.5
	(b)	$P(X<z)=0.10$ and $z=-1.28$
	(c)	$z=2.58$
	(d)	$z=1.0$
	(e)	$P(Z<z)=0.8+0.10749=0.90749$ and $z=1.33$.
3.71		99.73%
3.72	*(a)*	0.8849
	(b)	0.1151
	(c)	0.7262
3.73		0.8415
3.74		$x=9.771$
3.75		$P(X\leq 1)=0.923$ (Hypergeometric); $P(X\leq 1)=0.914$ (Binomial)
3.76		$P(X\leq 3)=0.429$; $P(X\leq 3)=0.433$ (Poisson)
3.77		$P(12\leq X\leq 14)=0.3716$
3.78		0.7748
3.79		0.755
3.80	*(a)*	0.1937
	(b)	0.1839

3.81 0.7717 (Hypergeometric); 0.7763 (Binomial)

3.82 $P(X = 7 \text{ or } 8) = 0.1611$

3.83 *(a)* 0.7286

(b) 0.6318

3.84 0.1131

3.85 *(a)* 0.0484

(b) 0.0192

(c) 0.1525

3.86 $P(X \leq 491.1 \text{ or } X \geq 50.9) \leq 1/4$

3.87 88.8%

3.89 75%

3.90 *(a)* 75%

(b) 84%

(c) 89%

3.91 *(a)* $b = 26.83$

(b) $\sigma = 1.58$

3.92 95%

3.93 0.0062

3.94 0.0136

3.95 0.7698

3.96 1/150

3.97 0.0808

3.98 Mean = 3.02; Standard deviation = 0.042

CHAPTER–4

4.1 *(a)*

Table 4.1: Sampling distribution of $\bar{x}$

Sample	Mean $\bar{x}$	Probability
1, 1	1	1/9
1, 2	1.5	1/9
1, 5	3	1/9
2, 1	1.5	1/9
2, 2	2	1/9
2, 5	3.5	1/9
5, 1	3	1/9
5, 2	3.5	1/9
5, 5	5	1/9

(b) The mean is $\mu = 8/3$.

(c) The mean of the sampling distribution of proportion is 8/3, and the population mean is also 8/3. In general, the distribution of sample means will have a mean equal to the population

mean. The sample means tend to target the population mean instead of systematically being too low or too high.

4.2 $P(\overline{X} < 95) = P(z < -2.5) = 0.006210$

4.3 The distribution of X and $\overline{X}$ are shown in Fig. 4.3.

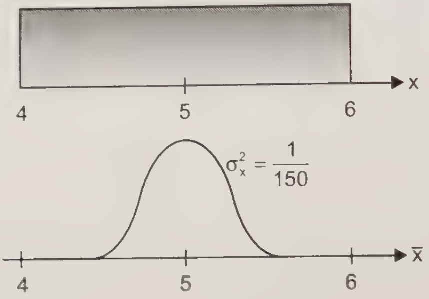


Fig. 4.3 The distribution of *x* and $\bar{x}$

4.4 *(a)* $\mu_{\bar{x}} = \mu = 60\,\text{kg} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{25}} = 0.6\,\text{kg}$

(b) $\mu_{\bar{x}} = \mu = 60\,\text{kg} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{3}{\sqrt{25}}\sqrt{\frac{4000-25}{4000-1}} = 0.5982$

4.5 *(a)* 0.96407

(b) 0.86218

4.6 *(a)* 67

(b) –2.67

4.7 Approximately 83.32% of the samples will have the mean between 588 and 605 hours.

4.8 *(a)* 0.21644

(b) 0.00154

4.9 *(a)* $z = \frac{\overline{X} - 55}{0.2}$ is $N(0, 1)$

(b) 0.9545

4.10 *(a)* 0.01224

(b) 0.88814

(c) 0.78870

(d) 0.006210

4.11 0.11507

4.12 *(a)* 0.02 mm

(b) $z = \frac{X - 12}{0.02}$ is $N(0, 1)$

(i) 0.30854 *(ii)* 0.158655 *(iii)* 0.53281

4.13 *(a)* 0.039204

(b) 0.710856

(c) 0.173609

4.14 There is a 96.25% probability that the estimate will be within 6 months of the true mean.

4.15 *(a)* $\sigma_{\bar{x}} = \frac{100}{\sqrt{100}} = 10$

(b) The probability is approximately 98.76% that the sampling error made in estimating the population means length of time employed by all men with one job by the mean length of time employed of a sample of 100 such men will be at most 25 weeks.

4.16 *(a)* $\mu = 80$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{16}{\sqrt{64}} = 2$

(b) $z = \frac{\bar{X} - 80}{2}$ is approximately $N(0, 1)$

(i) 0.02275 *(ii)* 0.00135 *(iii)* 0.5

4.17 *(a)* If all possible sample means for sample size 81 were found, their distribution would be approximately normal with mean \$45,000 and standard deviation \$1000.

(b) If all possible sample mean for samples of size 289 were found, their distribution would be approximately normal with mean \$45,000 and standard deviation \$529.42.

(c) The answer is no because these results follow from the central limit theorem (Chapter 3, Section 3.8), which implies that for large samples ($n \geq 30$), the distribution of the sample mean will be approximately normal regardless of the distribution of the original variable.

(d) 68.27% of samples of size 81 will have a mean that is within \$1000 of the population mean, i.e., 68.27% of the samples will have a sampling error less than \$1000 for samples of size 81.

(e) 94.12% of samples of size 289 will have a mean that is within \$1000 of the population mean, i.e., 94.12% of the samples will have a sampling error less than \$1000 for samples of size 289.

4.18 0.621%

4.19 *(a)* $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{9}{\sqrt{100}} = 0.9$

(b) The answer is it does not invalidate the result found in part (a). The sample size of 100 is large enough so that the central limit theorem (Chapter 3, Section 3.8) applies and the sampling distribution will be approximately normal.

(c) 0.99913

4.20 0.09428

4.21 $np = 49(0.20) = 9.8$ and $nq = 49(0.8) = 39.2$. Since np and nq are both greater than 5, the sampling distribution of $\hat{p}$ is approximately normal.

4.22 0.0914

4.23 0.9463

CHAPTER–5

5.1 3514.67 to 3617.33

5.2 49,000 to 50,196

5.3 (a) $n \approx 25$

(b) (i) $n \approx 35$ (ii) $n \approx 49$

5.4 $20.45 \le \mu \le 21.55$

5.5 $3.65 \le \mu \le 4.35$

5.6 $19.3156 \le \mu \le 20.6844$

5.7 $70.06 \le \mu \le 75.94$

5.8 $0.5089 \le \mu \le 0.5091$

5.9 $$\bar{X} - \frac{11.76}{\sqrt{n}} \le \mu \le \bar{X} + \frac{11.76}{\sqrt{n}}$$

5.10 $53.458 \le \mu \le 58.342$

5.11 $19.2308 \le \mu \le 32.1692$

5.12 $8.667 \le \mu \le 16.267$

5.13 (a) 18.646 to 83.175

(b) 21.527 to 66.692

(c) 23.239 to 59.886

5.14 $1.325 \le \sigma^2 \le 9.333$

5.15 $\sigma^2 \le 24.7875$

5.16 (a) 0.163

(b) (i) $0.0635 < \sigma^2 < 0.9807$ (ii) $0.2324 < \sigma^2 < 1.2129$

5.17 $0.9674 \le \sigma \le 1.9327$

5.18 $1.876 \le \sigma \le 5.833$

5.19 (a) $0.121 \le p \le 0.179$

(b) $0.141 \le p \le 0.159$

(c) $0.147 \le p \le 0.153$

5.20 0.3697 to 0.4303

5.21 (a) 0.669145 to 0.73085

(b) 0.66334 to 0.73665

(c) 0.651754 to 0.74824

(d) Yes

5.22 (a) $\hat{p} = 0.90; \pm 0.0294$

(b) 0.8706 to 0.9294

5.23 (a) 66.82% to 93.18%

(b) lowering the confidence level and increasing the sample size

5.24 0.57853 to 0.621471

5.25 12.16 to 17.84 years

5.26		5.2% to 24.8%
5.27		$0.61 \le \mu_1 - \mu_2 \le 35.81$
5.28		$2.593 < \mu_1 - \mu_2 < 5.407$
5.29		0.60 to 2.4 policies
5.30		–0.31 to 0.07
5.31	*(a)*	$-1.923 \le \mu_1 - \mu_2 \le 0.9225$
	(b)	Confidence interval is broadened
5.32		$46.282 \le \mu_1 - \mu_2 \le 53.718$
5.33		146.2 to 253.8 g
5.34		$-4.342 \le \mu_1 - \mu_2 \le 0.342$
5.35		$-0.72 \le \mu_1 - \mu_2 \le 6.72$
5.36		$-3.6759 < \mu_1 - \mu_2 < 5.6759$
5.37		–4.7428 to 15.314
5.38		–18.866 to –2.634
5.39		0.9396 to 18.4732
5.40		–7.2017 to –2.7983
5.41		1.2493 to 3.4174
5.42		–1.0332 to –5.5668
5.43		$0.1395 \le \frac{\sigma_A^2}{\sigma_B^2} \le 1.96$
5.44		$0.3244 \le \frac{\sigma_1^2}{\sigma_2^2} \le 2.58285$
5.45		$0.07592 \le \frac{\sigma_1^2}{\sigma_2^2} \le 2.86224$
5.46		$0.609375 \le \frac{\sigma_1^2}{\sigma_2^2} \le 4.453125$
5.47		$0.45 \le \frac{\sigma_1^2}{\sigma_2^2} \le 4.05$
5.48	*(a)*	$0.309 \le \frac{\sigma_1^2}{\sigma_2^2} \le 1.7475$
	(b)	$0.2565 \le \frac{\sigma_1^2}{\sigma_2^2} \le 2.115$
	(c)	$0.6124 \le \frac{\sigma_1^2}{\sigma_2^2}$

5.49	*(a)*	-0.2334
	(b)	$-0.293582 \le p_1 - p_2 \le -0.173218$
5.50	*(a)*	0.025
	(b)	-0.031626 to 0.081626
5.51		$-0.065387 \le p_1 - p_2 \le 0.109787$
5.52		$0.1213 \le p_1 - p_2 \le 0.4797$
5.53		-0.122774 to -0.070752
5.54		-0.113897 to -0.046103
5.55	*(a)*	$n \approx 865$
	(b)	$n \approx 385$
5.56	*(a)*	$n \approx 96$
	(b)	$n \approx 136$
5.57		$n \approx 35$
5.58		$n \approx 384$
5.59		$n \approx 32{,}617$
5.60	*(a)*	$n \approx 999$
	(b)	$n \approx 1040$
5.61		$n \approx 1849$
5.62		$n \approx 2122$
5.63	*(a)*	$n \approx 96{,}040$
	(b)	$n \approx 2167$

CHAPTER–6

6.1	*(a)*	± 2.58
	(b)	-2.33
	(c)	2.33
6.2	*(a)*	± 1.96
	(b)	-1.645
	(c)	1.645
6.3	*(a)*	± 2.81
	(b)	-2.58
	(c)	2.58
6.4	*(a)*	± 3.08
	(b)	-2.88
	(c)	2.88
6.5	*(a)*	The hypotheses that needed to tested: $H_0: \mu = 75$ $H_1: \mu > 75$
	(b)	Do not reject H_0

6.6 Reject H_0

6.7 $N = 139$

6.8 Sample mean cannot be taken as population means.

6.9 *(a)* Yes; H_0 is accepted; man time can be taken as 58.2 minutes
  (b) Do not reject H_0

6.10 Reject H_0; claim of the earlier study is not true.

6.11 Reject H_0; mean population time for the product is more than 45 minutes.

6.12 *(a)* Do not reject H_0
  (b) *(i)* $n = 65$ *(ii)* $n = 81$

6.13 Reject H_0

6.14 Reject H_0; mean number of machine 1 is higher than that by machine 2.

6.15 Reject H_0; the two brands of machine parts differ in their carbon content.

6.16 *(a)* reject H_0
  (b) do not reject H_0
  (c) reject H_0

6.17 Reject H_0; mean load at failure exceeds 10 MPa.

6.18 Reject H_0; mean baking temperature exceeds 90°C.

6.19 Reject H_0; mean weekly sales of salespersons increase as a result of attending that course.

6.20 Do not reject H_0; mean of the paired differences is not different from zero.

6.21 Reject H_0; use of gasoline additive increases the gasoline mileage.

6.22 *(a)* 0.85 to 17.15
  (b) Reject H_0; the mean number of items produced per hour by the first machine is higher than by the second machine.

6.23 Reject H_0.

6.24 Reject H_0.

6.25 Reject H_0.

6.26 Reject H_0.

6.27 Reject H_0; persons with a business degree are better salespersons than those with a non-business degree.

6.28 Do not reject H_0; the mean GPAs of all male and female college students are not different.

6.29 Reject H_0 and conclude that σ is not equal to 0.25 at $\alpha = 0.05$.

6.30 Reject H_0; the process used in the manufacture of parts must be adjusted.

6.31 Do not reject H_0.

6.32 Do not reject H_0; variance of current seniors is not different from 150.

6.33 Reject H_0; the population variance is greater than 0.003 mm.

6.34 Do not reject H_0; σ_2 is equal to 4000 hours.

6.35 Do not reject H_0; sufficient evidence to warrant that $\sigma \neq 10$.

6.36 Do not reject H_0; both production lines have the same variance from the predetermined length.

6.37 *(a)* Do not reject H_0; variances are equal.

(b) Do not reject H_0; no evidence to indicate that the new voltage-regulating device has affected the mean value.

6.38 Do not reject H_0; no strong evidence to indicate that either gasket material results in a smaller variance of gasket thickness.

6.39 Accept H_0; there is no significant difference between the variances.

6.40 Reject H_0.

6.41 Do not reject H_0.

6.42 Do not reject H_0; we cannot claim that the standard deviations of surface roughness for the two processes are different at the 90% confidence level.

6.43 Reject H_0.

6.44 Sufficient evidence to warrant rejection of the claim, the proportion of 0.60.

6.45 Reject H_0.

6.46 Do not reject H_0.

6.47 Do not reject H_0; no evidence that the proportion of defective bolts produced by the machine is reduced.

6.48 Reject H_0.

6.49 Do not reject H_0; no evidence to indicate that the attitude this year is different from that of last year.

6.50 Do not reject H_0; no strong evidence to support the claim that the two polishing materials are different.

6.51 Do not eject H_0; the data provides insufficient evidence to state that the fraction defective differ at $\alpha = 0.05$.

6.52 *(a)* 0.10

(b) 0.4382 to 0.15618

(c) The non-rejection region lies to the left of $z = 1.96$.

(d) $\hat{p} = 0.74545$; $\hat{q} = 0.254545$

(e) Reject H_0

6.53 Do not reject H_0.

6.54 Accept H_0

6.55 Do not reject H_0; no evidence to indicate that one brand of spray is significantly different from the other.

6.56 Reject H_0; there is sufficient evidence to conclude that the percentage of men who sometimes order vegetarian meals is smaller than the percentage of women who sometimes order vegetarian meals.

CHAPTER–7

7.1 $a = -22.0007, b = 3.7484$

7.2 $a = -149.3849, b = 9.1956$

7.3 $a = 0.4449, b = 0.2747$

7.4 $a = -47.4598, b = 11.582$

7.5 $a = 22.917, b = 3.3883$
7.6 $a = 44.3333, b = 0.4167$
7.7 $a = 80.7777, b = 1.138$
7.8 $a = 39.3327, b = 0.4213$
7.9 $a = 1.1574, b = 0.2566$
7.10 $a = 124.0862, b = -13.5402$
7.11 *(a)* $a = -22.0007$ and $b = 3.7484$
(b) $s_e = 30.8722$
(c) SSE = 7624.7219
(d) SST = 23973.6
(e) SSR = 16348.8781
(f) $r^2 = 0.682$
(g) $r = 0.8258$
when $x = 42, y = 135.4307$
7.12 *(a)* $a = -149.3849$ and $b = 9.1956$
(b) $s_e = 58.8065$
(c) SSE = 13832.8076
(d) SST = 125520.8333
(e) SSR = 111688.0258
(f) $r^2 = 0.8898$
(g) $r = 0.9433$
when $x = 53, y = 337.9811$
7.13 *(a)* $a = 0.4449$ and $b = 0.2747$
(b) $s_e = 1.0550$
(c) SSE = 5.5648
(d) SST = 66
(e) SSR = 60.4352
(f) $r^2 = 0.9157$
(g) $r = 0.9569$
when $x = 42, y = 11.9825$
7.14 *(a)* $a = -47.4598$ and $b = 11.582$
(b) $s_e = 5.5803$
(c) SSE = 249.1216
(d) SST = 47682.1
(e) SSR = 47432.9784
(f) $r^2 = 0.9948$
(g) $r = 0.9974$
when $x = 20, y = 184.1804$
7.15 *(a)* $a = 22.917$ and $b = 3.3883$

(b) $s_e = 5.6669$

(c) $\text{SSE} = 256.9085$

(d) $\text{SST} = 4573.6$

(e) $\text{SSR} = 4316.6915$

(f) $r^2 = 0.9438$

(g) $r = 0.9715$

when $x = 15, y = 73.7415$

7.16 *(a)* $a = 44.3333$ and $b = 0.4167$

(b) $s_e = 12.6573$

(c) $\text{SSE} = 1281.6667$

(d) $\text{SST} = 1802.5$

(e) $\text{SSR} = 520.8333$

(f) $r^2 = 0.2890$

(g) $r = 0.5375$

when $x = 82, y = 78.5$

7.17 *(a)* $a = 80.7777$ and $b = 1.138$

(b) $s_e = 7.0176$

(c) $\text{SSE} = 492.4669$

(d) $\text{SST} = 2500.6667$

(e) $\text{SSR} = 2008.1998$

(f) $r^2 = 0.8031$

(g) $r = 0.8961$

when $x = 45, y = 131.988$

7.18 *(a)* $a = 39.3327$ and $b = 0.4213$

(b) $s_e = 1.3376$

(c) $\text{SSE} = 17.8917$

(d) $\text{SST} = 32.9167$

(e) $\text{SSR} = 15.0249$

(f) $r^2 = 0.4565$

(g) $r = 0.6756$

when $x = 72, y = 69.6634$

7.19 *(a)* $a = 1.1574$ and $b = 0.2556$

(b) $s_e = 1.0438$

(c) $\text{SSE} = 5.4473$

(d) $\text{SST} = 58.2143$

(e) $\text{SSR} = 52.767$

(f) $r^2 = 0.9064$

(g) $r = 0.9521$

when $x = 41, y = 11.6778$

7.20 (*a*) $a = 124.0862$ and $b = -13.5402$

(*b*) $s_e = 15.4208$

(*c*) SSE = 1426.8046

(*d*) SST = 9402

(*e*) SSR = 7975.1954

(*f*) $r^2 = 0.85$

(*g*) $r = 0.92$

when $7 = 53$, $y = 29.3046$

7.21 1.6614 to 5.8354

7.22 4.7038 to 13.6874

7.23 0.1789 to 0.3705

7.24 10.8977 to 12.2663

7.25 2.7144 to 4.0622

7.26 –0.1162 to 0.9496

7.27 0.7410 to 1.5351

7.28 0.0974 to 0.7451

7.29 0.1618 to 0.3514

7.30 –19.2615 to –7.8189

7.31 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 2.896$; test statistic: $t = 4.1419$; reject H_0.

7.32 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 3.747$; test statistic: $t = 5.683$; reject H_0.

7.33 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 3.365$; test statistic: $t = 7.369$; reject H_0.

7.34 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 2.896$; test statistic: $t = 39.0283$; reject H_0.

7.35 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 2.896$; test statistic: $t = 11.5939$; reject H_0.

7.36 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 2.896$; test statistic: $t = 1.8030$; do not reject H_0.

7.37 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 2.764$; test statistic: $t = 6.3858$; reject H_0.

7.38 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 2.764$; test statistic: $t = 2.8979$; reject H_0.

7.39 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 3.365$; test statistic: $t = 6.9595$; reject H_0.

7.40 H_0: $B = 0$; H_1: $B > 0$; critical value: $t = 3.143$; test statistic: $t = -5.7911$; do not reject H_0.

7.41 (*a*) 2.0650 to 5.4317

(*b*) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 1.86$; test statistic; $t = 4.1419$; do not reject null hypothesis and B is positive.

7.42 (*a*) 5.7458 to 12.6453

(*b*) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 2.132$; test statistic; $t = 5.6829$; do not reject null hypothesis and B is positive.

7.43 (*a*) 0.1996 to 0.3498

(*b*) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 2.015$; test statistic; $t = 7.3646$; do not reject null hypothesis and B is positive.

7.44 (*a*) 11.03 to 12.134

(*b*) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 1.86$; test statistic; $t = 39.0229$; do not reject null hypothesis and B is positive.

7.45 *(a)* 2.8447 to 3.9319
(b) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 1.86$; test statistic; $t = 11.5958$; do not reject null hypothesis and B is positive.

7.46 *(a)* –0.0132 to 0.8465
(b) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 1.86$; test statistic; $t = 1.8031$; do not reject null hypothesis and B is positive.

7.47 *(a)* 0.8151 to 1.4609
(b) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 1.812$; test statistic; $t = 6.3861$; do not reject null hypothesis and B is positive.

7.48 *(a)* 0.1579 to 0.6847
(b) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 1.812$; test statistic; $t = 2.8975$; do not reject null hypothesis and B is positive.

7.49 *(a)* 0.1823 to 0.3309
(b) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 2.015$; test statistic; $t = 6.9539$; do not reject null hypothesis and B is positive.

7.50 *(a)* –18.0831 to 8.9973
(b) H_0: $B = 0$; H_1: $B < 0$; critical value of $t = 1.942$; test statistic; $t = -5.7911$; reject the null hypothesis and B is positive.

7.51 86.8864 to 183.9750

7.52 196.601 to 479.3612

7.53 9.6917 to 14.2733

7.54 177.8678 to 190.4931

7.55 65.9838 to 81.4991

7.56 62.1632 to 94.8368

7.57 124.3482 to 139.6277

7.58 66.9186 to 70.4082

7.59 9.4268 to 13.9288

7.60 4.0352 to 54.5739

7.61 $b = 0.8671$ and $m = 1.3471$

7.62 $b = 0.5686$ and $m = 1.5903$

7.63 $b = 0.3990$ and $m = 0.9057$

7.64 $b = 2.9256$ and $m = 1.3777$

7.65 $b = 18.9926$ and $m = 0.4983$

7.66 $a = 1.2675$, $b = 1.2475$ and $c = -0.0727$; $S_{y/x} = 0.6520$; $r = 0.9653$

7.67 $a = -291.9285$, $b = 417.3647$ and $c = -146.3105$; $S_t = 8.2232$; $S_r = 0.0406$; $r = 0.9975$; $S_{y/x} = 0.1163$

7.68 $a = 24.1204$, $b = 1.9639$, $c = -0.0323$, $S_{y/x} = 1.5211$, $S_t = 288$, $S_r = 4.6277$, $r^2 = 0.9839$, $r = 0.9919$

7.69 $a = 2.5095$, $b = 1.2$, $c = 0.7333$, $S_{y/x} = 0.2024$, $S_t = 332.0571$, $S_r = 0.1638$, $r^2 = 0.9995$, $r = 0.9998$

7.70 $a = 18.9414$, $b = -0.0313$, $c = 0.0014$, $S_{y/x} = 4.7437$, $S_t = 211.755$, $S_r = 67.5086$, $r^2 = 0.6812$, $r = 0.8253$

7.71 21.043 to 249.8185
7.72 32.545 to 643.4171
7.73 7.1513 to 16.8138
7.74 164.4228 to 203.9381
7.75 53.2073 to 94.2757
7.76 33.0006 to 123.9994
7.77 108.4735 to 155.5024
7.78 64.6135 to 74.7133
7.79 6.9051 to 16.4505
7.80 –33.1963 to 91.8055
7.81 $a = 14.4217$, $b = 8.9904$, $c = -5.6087$, $S_t = 1051.3356$, $S_r = 4.3466$, $S_{y/x} = 0.7244$, $r^2 = 0.9959$, $r = 0.9979$
7.82 $a = 7.3427$, $b = -4.5932$, $c = -2.7062$, $S_t = 728.5556$, $S_r = 50.742$, $S_{y/x} = 8.457$, $r^2 = 0.9304$, $r = 0.9645$
7.83 $a = 2.2561$, $b = 3.8171$, $c = 5.3333$, $S_t = 577.84$, $S_r = 4.2007$, $S_{y/x} = 0.6001$, $r^2 = 0.9927$, $r = 0.9964$
7.84 $a = 23.9103$, $b = -4.691$, $c = -9.591$, $S_t = 850.36$, $S_r = 1.9791$, $S_{y/x} = 0.2827$, $r^2 = 0.9977$, $r = 0.9988$
7.85 $a = 29.4487$, $b = -5.7449$, $c = 19.8449$, $S_t = 3463.61$, $S_r = 2.0574$, $S_{y/x} = 0.2939$, $r^2 = 0.9994$, $r = 0.9997$
7.86 $a = 107.6961$ and $b = -0.0143$
7.87 $a = 2.7077$ and $b = 0.0369$
7.88 $a = 20.0771$ and $b = 0.0325$
7.89 $a = 63.5770$ and $b = 0.3548$
7.90 $a = 1.9077$ and $b = 1.5146$
7.91 $a = 4.0672$ and $b = 0.7823$
7.92 $a = 2.3946$ and $b = 177.5202$
7.93 $a = 13.2275$ and $b = 37.5036$
7.94 $a = 1.8515$ and $b = 284.4318$
7.95 $a = -0.0359$ and $b = 7.6495$; $r = 0.9998$
7.96 $a = 25.0550$; $b = 3.2229$; Sum of squares of errors = 269.5377
7.97 $a = 3.3385$; $b = 2.9169$; Sum of squares of errors = 7.8421
7.98 $a = -88.6813$; $b = 7.4678$; Sum of squares of errors = 205.2290
7.99 $a = 1.2045$; $b = 0.2599$; Sum of squares of errors = 1.6846
7.100 $a = -40.8141$; $b = 11.1799$; Sum of squares of errors = 308.9377
7.101 $a = 21.5949$; $b = 3.5681$; Sum of squares of errors = 271.3282
7.102 $a = 22.4556$; $b = 3.13$; Sum of squares of errors = 86.0916
7.103 $a = 5.2116$; $b = 2.8682$; Sum of squares of errors = 5.6054
7.104 $a = -88.3833$; $b = 7.5646$; Sum of squares of errors = 17572.5447
7.105 $a = -1.2576$; $b = 0.3166$; Sum of squares of errors = 1.1108
7.106 $a = -49.1549$; $b = 11.7656$; Sum of squares of errors = 41.9142
7.107 $a = 18.6429$; $b = 3.4286$; Sum of squares of errors = 14.1888

7.108 $y = 1.01299 + 0.85113x + 0.83918x^2$

7.112 $\phi_0(x) = 1$, $\phi_1(x) = x - 0.5$, $\phi_2(x) = x^2 - x + \frac{1}{6}$, and $\phi_3(x) = x^3 - 1.5x^2 + 0.6x - 0.05$.

7.113 $\phi_0(x) = 1$, $\phi_1(x) = x - 1$, $\phi_2(x) = x^2 - 2x + \frac{2}{3}$, and $\phi_3(x) = x^3 - 3x^2 + \frac{12}{5}x - \frac{2}{5}$

7.114 $\phi_0(x) = 1$, $\phi_1(x) = x - 2$, $\phi_2(x) = x^2 - 4x + \frac{11}{3}$, and $\phi_3(x) = x^3 - 6x^2 + 11.4x - 6.8$

CHAPTER–8

8.1 Critical value: $\chi^2 = 11.345$; test statistic: $\chi^2 = 17.6$; reject H_0.

8.2 Critical value: $\chi^2 = 5.991$; test statistic: $\chi^2 = 1.062$; do not reject H_0.

8.3 Critical value: $\chi^2 = 13.277$; test statistic: $\chi^2 = 16.3$; reject H_0.

8.4 Critical value: $\chi^2 = 7.378$; test statistic: $\chi^2 = 9.158$; reject H_0.

8.5 Critical value: $\chi^2 = 11.071$; test statistic: $\chi^2 = 2.48$; do not reject H_0.

8.6 Expected frequencies: 105, 59.5, 10.5, 75, 42.5 and 7.5.

8.7 Critical value: $\chi^2 = 9.21$; test statistic: $\chi^2 = 5.2$; do not reject H_0.

8.8 Critical value: $\chi^2 = 5.991$; test statistic: $\chi^2 = 199.955$; reject H_0.

8.9 Critical value: $\chi^2 = 9.21$; test statistic: $\chi^2 = 63.0848$; reject H_0.

8.10 Critical value: $\chi^2 = 7.378$; test statistic: $\chi^2 = 7.7749$; reject H_0.

8.11 Critical value: $\chi^2 = 7.378$; test statistic: $\chi^2 = 5.239$; do not reject H_0.

8.12 Critical value: $\chi^2 = 4.605$; test statistic: $\chi^2 = 14.5939$; reject H_0.

8.13 Critical value: $\chi^2 = 16.812$; test statistic: $\chi^2 = 10.1218$; do not reject H_0.

8.14 Critical value: $\chi^2 = 6.635$; test statistic: $\chi^2 = 15.1294$; reject H_0.

8.15 Critical value: $\chi^2 = 7.378$; test statistic: $\chi^2 = 1.9911$; do not reject H_0.

8.16 0.01829 to 0.05805; 0.13524 to 0.2409.

8.17 Test statistic: $\chi^2 = 22.2$; do not reject H_0.

8.18 Test statistic: $\chi^2 = 26.565$; reject H_0.

8.19 Test statistic: $\chi^2 = 64.96$; reject H_0.

8.20 Test statistic: $\chi^2 = 25.6$; do not reject H_0.

8.21 0.003364 to 0.011972; 0.058 to 0.10942

8.22 2528.589 to 11652.842; 50.285 to 107.948

8.23 Critical value: $\chi^2 = 42.98$; test statistic: $\chi^2 = 48$; reject H_0.

8.24 Critical value: $\chi^2 = (32.852, 8.907)$; test statistic: $\chi^2 = 21.913$; do not reject H_0.

CHAPTER–9

9.1 $F = 1.2742$

9.2 Test statistic $F = 1.2742$ and $F_{\text{critical}} = 6.93$. We fail to reject the null hypothesis. The means of three populations are equal.

9.3 Test statistic $F = 10.72 >$ the critical value of $F = 3.16$. We reject the null hypothesis.

9.4 Critical of $F = 4.75$ and test statistic $F = 2.4291$. Do not reject H_0.

9.5 Critical of $F = 3.47$ and test statistic $F = 20.6516$. Reject the null hypothesis.

9.6 Critical of $F = 6.93$ and test statistic $F = 1.5164$. Do not reject H_0

9.7 Critical of $F = 3.59$ and test statistic $F = 5.8068$. Reject H_0

9.8 Critical of $F = 3.34$ and test statistic $F = 7.0634$. Reject H_0

9.9 Critical of $F = 3.24$ and test statistic $F = 9.3653$. Reject H_0

9.10 Critical of $F = 3.49$ and test statistic $F = 6.6758$. Reject H_0

YEOVIL COLLEGE
LIBRARY

YEOVIL COLLEGE
LIBRARY

PROBABILITY AND STATISTICS
FOR SCIENTISTS AND ENGINEERS

Contents:

- Numerical Descriptive Measures
- Probability
- Random Variables and Probability Distributions
- Sampling Distributions
- Estimation
- Hypothesis Testing
- Curve Fitting, Regression and Correlation
- Chi-Square Tests
- Analysis of Variance
- Appendices

Key Features:

- Coverage of all course fundamentals in easy-to-understand methodology.
- Clarity in the presentation of concepts.
- Review questions, true/false, and fill in the blanks for each chapter.
- Over 230 fully solved problems with step-by-step solutions.
- Over 520 additional practice problems with answers.

About the Book:

Probability and Statistics for Scientists and Engineers is designed as a supplement to an introductory undergraduate or graduate course for mathematics, science and engineering students of all disciplines. The text covers all major aspects of engineering statistics, including descriptive statistics, probability and probability distributions, statistical test and confidence intervals for one and two samples and curve fitting, correlation, regression, chi-square tests, and analysis of variance. The book continues to maintain a student-friendly approach and numerical problem solving orientation. Presentations are limited to very basic topics to serve as an introduction to advance topics in those areas of discipline. The purpose of the book is to present the principles and concepts of Probability and Statistics as relevant to student learning.

About the Author:

Dr Rao V Dukkipati is Professor, Chair and Director of the Graduate Programme of the Department of Mechanical Engineering at Fairfield University, Connecticut, USA. He also serves as an adjunct professor in the Mechanical Engineering Department at Concordia University, Montreal, Canada. He gained his doctorate from Oklahoma State University, Stillwater, USA in 1973. His specialized areas of teaching and research include mechanical design, mechanics, vibrations, control systems, economy and reliability, quality control improvement and management, vehicle dynamics, and accident reconstruction. He has to his credit over 300 technical papers and 50 books, and is a fellow of ASME and CSME.

Readership & Market Potential: UG/PG students and teachers of mathematics, science and engineering • University and college libraries • Design and research engineers • Professionals and practising engineers.

ISBN: 978-1-906574-83-3

KS-097-431

New Academic Science Limited
The Control Centre, 11 A Little Mount Sion
Tunbridge Wells, Kent TN1 1YS, UK
www.newacademicscience.co.uk
e-mail: info@newacademicscience.co.uk